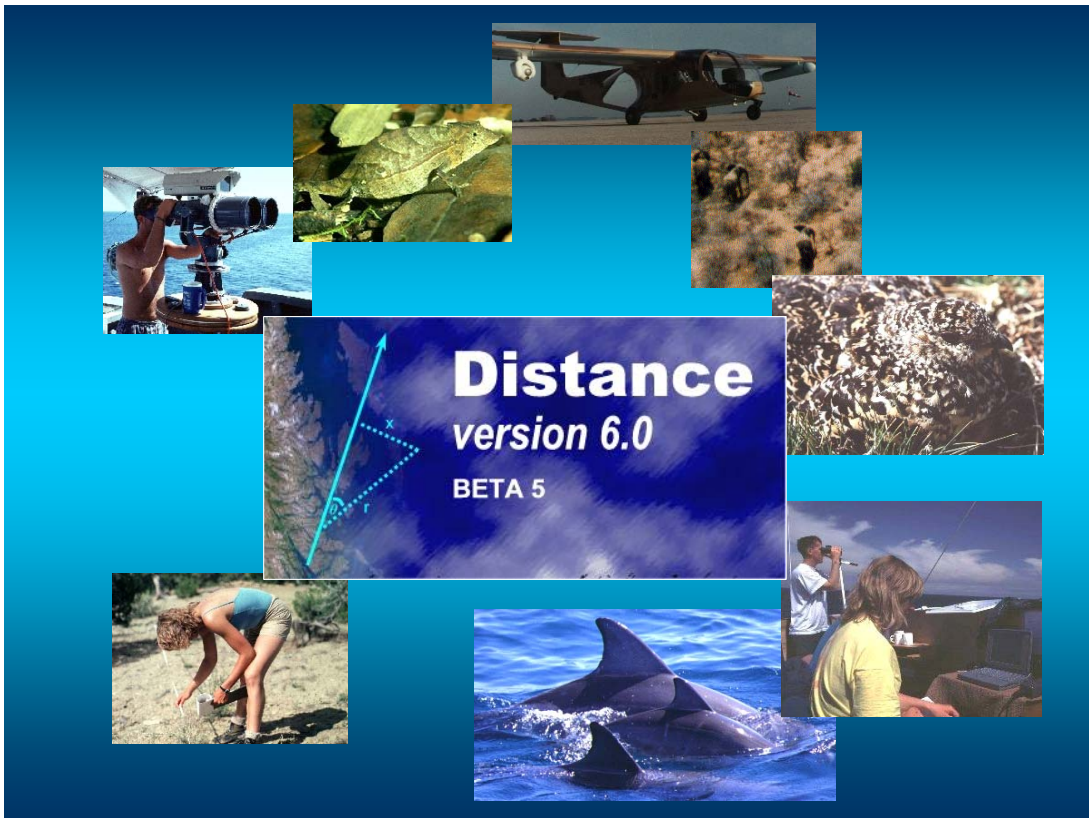
User's Guide

Distance 6.0 Beta 5

Research Unit for Wildlife Population Assessment



Contents

Chapter 1 - Introduction 1

Welcome 1
- Conventions in This Documentation 2
- Where to Go Next 2
- Citation for Distance 3
- Distance Sampling Reference Books 3

Staying in Touch 3
- Distance-sampling Email List 3
- Program Distance Web Site 4
- Sending Suggestions and Reporting Problems 4

Chapter 2 - About Distance 5

What is Distance? 5
Use Agreement 5
Sponsors 5
Distance Development Team 6
Acknowledgements 6
History of Distance 7
New in Distance 8
- New Features of Distance 6.0 8
- New Features of Distance 5.0 8
- New Features of Distance 4.1 8
- New Features of Distance 4.0 8
- New Features of Distance 3.5 10
- New Features Planned for Future Versions 12

Chapter 3 - Getting Started 13

Objective 13
Example 1: Using Distance to Analyze Simple Data 13
- Example 1 - Preparing the data for import 13
- Example 1 - Creating the Distance project 14
- Example 1 - Importing the data 15
- Example 1 - Studying the data in Distance 16
- Example 1 – Running the first analysis 16
- Example 1 - Creating a new analysis 17
- Example 1 - Further investigations 18

Example 2: More Complex Data Import 19
- Example 2 - Preparing to import the data 19
- Example 2 - Importing the data 20
- Example 2 - Analysis 20

Example 3: Using Distance to Design a Survey 21
- Example 3 - Preparing the data for import 21
- Example 3 - Creating the Distance project 22
- Example 3 - Importing the Geographic Data 22
- Example 3 - Checking the Data on a Map 22
- Example 3 - Adding a Coverage Layer 23
- Example 3 - Creating a Survey Design 24
- Example 3 - Automated Generation of New Surveys 24

Example 3 - Design Statistics 25
Example 3 - Further Investigations 26
Example 4 - A Second Survey Design Project 26
Example 4 - Opening the Mexico Project in Distance 26
Example 4 - Reviewing the Project Properties 26
Example 4 - Examining the Data 27
Example 4 - Creating a New Design 28
Example 4 - Automated Generation of New Surveys 29
Example 4 - Design Statistics 31
Example 4 - Further Investigations 31
Sample Projects 31

Chapter 4 - Distance Projects 33

Introduction to Distance Projects 33
Creating a New Project 34
Using an Existing Project as a Template 35
Opening an Existing Project 36
Saving and Backing Up a Project 36
Saving Projects 36
Backing up Projects 36
Exporting, Transporting and Archiving Projects 38
Viewing and Editing Project Properties 39
Compacting a Project 39
Importing from Previous Versions of Distance 39
Importing Distance 4 and 5 Projects 39
Importing Distance 3.5 Projects 40
Importing Distance 2.2 - 3.0 Command Files 40

Chapter 5 - Data in Distance 41

Data Structure 41
Data Layers 41
Data Fields 43
Changing the Data Structure 45
Getting Data Into Distance 45
Data Entry 46
Data Import 46
Geographic (GIS) Data 52
Viewing and Manipulating Geographic Data 52
Coordinate Systems and Projections 53
GIS Data Format 56
Importing Existing GIS Data 57
Advanced Data Topics 61
Linking to Data From Other Databases 61

Chapter 6 - Survey Design in Distance 63

Introduction to Survey Design in Distance 63
Design Classes Available in Distance 64
Survey Design Concepts 65
Concept: Coverage Probability 65
Concept: Edge Effects 65
Concept: Zigzag Sampling Designs 66
Setting Up a New Project for Survey Design 68

Chapter 7 - Analysis in Distance 71

Introduction to Analysis in Distance 71
Introduction to the Analysis Browser 71
Introduction to Analysis Details Windows 72

Analysis Components 72
Data Filters, Model Definitions and Survey Objects 72
Working with Data Filters and Model Definitions 74
Working with Surveys during Analysis 78
Analysis Engines 80
Conventional Distance Sampling (CDS) Engine 80
Multiple Covariate Distance Sampling (MCDS) Engine 81
Mark Recapture Distance Sampling (MRDS) Engine 81
Density Surface Modelling (DSM) Engine 81
Running Analyses 81
Locking the Data Sheet 82
Cleaning the Temp folder 83
R Statistical Software 83
Installing and Configuring R 84
Updating the Version of R That Distance Uses 84
Contents of the R folder 84
Images produced by R 84

Chapter 8 - Conventional Distance Sampling Analysis 87

Introduction to CDS Analysis 87
Setting up a Project for CDS Analysis 88
CDS Analysis Guidelines 88
Output from CDS Analyses 89
CDS Results Details Listing 90
About CDS Detection Function Formulae 95
CDS Analysis Browser Results 98
Exporting CDS Results 98
Miscellaneous CDS Analysis Topics 100
Interval (Binned/Grouped) Data 100
Missing Data in CDS Analysis 101
Clusters of Objects 102
Stratification and Post-stratification 103
Variance Estimation in CDS 107
Multipliers in CDS Analysis 108
Model Averaging in CDS Analysis 111
Sample Definition in CDS Analysis 112
Unknown Study Area Size 112
Restricting Inference to Density or Abundance in the Covered Region in CDS Analysis 112
Analysis of Data from a Single Transect in CDS 113
Running CDS Analyses From Outside Distance 114

Chapter 9 - Multiple Covariates Distance Sampling Analysis 115

Introduction to MCDS Analysis 115
Introducing the MCDS Engine 115
Setting up a Project for MCDS Analysis 116
Defining MCDS Models 116
MCDS Analysis Guidelines 120
Choosing Covariates to Include in the Model 120
Specifying the Model 121
Truncation for MCDS Analyses 122
Output from MCDS Analyses 122
MCDS Results Details Listing 122
Exporting MCDS Results 124
Miscellaneous MCDS Analysis Topics 124
Missing Data in MCDS Analysis 124
Stratification and Post-stratification in MCDS 125
Running MCDS Analyses from Outside Distance 125
Analysis of Double Observer Data with the MCDS Engine 125

Chapter 10 - Mark Recapture Distance Sampling 127

Introduction to Mark Recapture Distance Sampling....127
Setting up a Project for MRDS Analysis....129
Setting up Your Data for MRDS Analysis....129
Defining MRDS Models....130
Introduction to MRDS Models....130
DS and MR Models....131
Specifying DS and MR Model Formulae....131
MRDS Analysis Guidelines....134
Output from MRDS Analyses....135
MRDS Results Details Listing....135
MRDS Analysis Browser Results....136
Exporting MRDS Results....136
Miscellaneous MRDS Analysis Topics....137
Interval Data in MRDS....137
Clusters of Objects in MRDS....137
Stratification and Post-stratification in MRDS....137
Variance Estimation in MRDS....137
Multipliers in MRDS Analysis....138
Model Averaging in MRDS Analysis....138
Sample Definition in MRDS Analysis....139
Using a Previously Fitted Detection Function to Estimate Density in MRDS....139
Restricting Inference to Density or Abundance in the Covered Region in MRDS Analysis....140
Running the MRDS Analysis Engine from Outside Distance....140
Installing an Updated Version of the MRDS Engine....141
Checking Which Version of the MRDS Engine is Being Used....141
Fine-tuning an MRDS Analysis....142
Single Observer Configuration in the MRDS Engine....143

Chapter 11 – Density Surface Modelling 145

Introduction to Density Surface Modelling....145
Setting up a Project for DSM Analysis....146
Setting up Your Data for DSM Analysis....147
Defining DSM Models....151
Introduction to DSM Models....151
Specifying DSM Model Formulae....151
DSM Analysis Guidelines....152
Density surface model....152
Prediction of abundance to unsurveyed areas....153
Variance estimation using parametric bootstrap....154
Output from DSM Analyses....154
DSM Results Details Listing....155
DSM Analysis Browser Results....156
Exporting DSM Results....156
Miscellaneous DSM Analysis Topics....156
Clusters of Objects in DSM....156
Stratification and Post-stratification in DSM....157
Running the DSM Analysis Engine from Outside Distance....157
Installing an Updated Version of the DSM Engine....158
Checking Which Version of the DSM Engine is Being Used....158
Fine-tuning a DSM Analysis....159

Chapter 12 - Troubleshooting 161

Known Problems....161
Internal Errors in the Interface....161
Problems with the Analysis Engines....162
Errors and Warnings in the CDS and MCDS Analysis Engines....162

Problems with the MRDS Engine 162
Stopping an Analysis 163
GIS Problems 163
Recovering from Unexpected Program Exit 164
Fixing a corrupted project 164

Appendix - Program Reference 165

Introduction to Program Reference 165
Setup Project Wizard 165
Setup for Analyzing a Survey 166
Setup for Designing Surveys 169
Use Another Project as Template 169
Import from Previous Version of Distance 169
Data Entry Wizard 170
Global Layer Wizard Page 171
Stratum Layer Wizard Page 171
Sample Layer Wizard Page 171
Observation Layer Wizard Page 171
Finished Data Entry Wizard Page 172
Import Data Wizard 172
Data Source Wizard Page 172
Data Destination Wizard Page 173
Data File Format Wizard Page 174
Data File Structure Wizard Page 174
Finished Import Data Wizard 176
Troubleshooting the Import Data Wizard 176
Project Properties Dialog 177
General Project Properties Tab 177
Geographic Project Properties Tab 178
Preferences Dialog 178
General Preferences Tab 178
Geographic Preferences Tab 180
Survey Design Preferences Tab 180
Analysis Preferences Tab 181
Project Browser 183
Data Explorer 183
Data Layers Viewer 185
Data Sheet 187
Map Browser 192
Design Browser 193
Survey Browser 195
Analysis Browser 197
Map Window 199
Design Details Window 201
Design Details Inputs Tab 201
Design Details Log Tab 202
Design Details Results Tab 203
Survey Details Window 209
Survey Details Inputs Tab 209
Survey Details Log Tab 210
Survey Details Results Tab 210
Analysis Details Window 210
Analysis Details Inputs Tab 210
Analysis Details Log Tab 213
Analysis Details Results Tab 214
Design Properties Dialog 215
General Design Properties Tab 216
Coverage Probability Design Properties Tab 217
Sampler Design Properties Tab 218

Effort Allocation Design Properties Tab 219
Survey Properties Dialog 230
Survey Methods Survey Properties Tab 230
Data Layers Survey Properties Tab 231
Data Fields Survey Properties Tab 231
Data Filter Properties Dialog 231
Data Selection Tab 232
Intervals Tab 233
Truncation Tab 235
Units Tab 236
Model Definition Properties Dialog 236
Model Definition Properties - CDS and MCDS 237
Model Definition Properties - MRDS 250
Model Definition Properties – DSM 253
Analysis Components Window 253
Other Windows 254
About Distance Dialog 254
Export Project Dialog 254
Projection Parameters Dialog 255
Create New Layer Dialog 255
Grid Properties Dialog 255
Insert or Append Field Dialog 256
Data Layer Properties Dialog 256
Shape Properties Dialog 257
New Coordinate System Dialog 257
Column Manager Dialog 258
Arrange Sets Dialog 258
Map Properties Dialog 258
Add Map Layer Dialog 258
Run Design Dialog 259
Confirm Change Dialog 259
R Image Properties Dialog 259
Data Selection Zoom Dialog 260

Appendix - Inside Distance 261

Introduction to Inside Distance Appendix 261
Distance components 261
The Distance 6 Database API 261
Data File Reference 262
How Distance Stores Data 262
Linking to External Data from Distdata.mdb 267
Valid Names 276
Miscellaneous topics 278
Random number generation 278

Appendix - MCDS Engine Reference 279

Introduction to MCDS Engine Reference 279
Some history 279
Running the MCDS engine 280
MCDS Command Language 281
Header Section 282
Options Section 283
Data section 291
Estimate section 293
MCDS Engine Required Data Format 308
Output From the MCDS Engine 309
MCDS Engine Command Line Output 310
MCDS Engine Output File 310
MCDS Engine Log File 311

MCDS Engine Stats File 311
MCDS Engine Plot File 312
MCDS Engine Bootstrap File 313
MCDS Engine Bootstrap Progress File 313
MCDS Engine Limitations 314
MCDS Engine Fitting Algorithms 314
MCDS Engine Error and Warning Messages 315
MCDS Engine Warning Messages 315
MCDS Engine Error Messages 319
MCDS Engine Internal Error Messages 323
Changes in MCDS Engine Since Distance 2.2 324

Appendix – HT estimation of density when probability of coverage is unequal 327

Review of the Horvitz-Thompson-like estimator 327
Setting up a Project for DHT Analysis 328
Output from DHT Analyses 329
DHT Results Details Listing 329
Exporting DHT Results 329
Miscellaneous DHT Analysis Topics 330
Clusters of Objects in DHT 330
Stratification and Post-stratification in DHT 330
Running the DHT Analysis Engine from Outside Distance 330
Checking Which Version of the DHT Engine is Being Used 331

Bibliography 333

Glossary of Terms 335

Index 343

Chapter 1 - Introduction

Welcome

This is a beta release of Distance 6.0, and the contents of the manual are still under construction!

Welcome to Distance 6.0!

Distance software allows you to design and analyze distance sampling surveys, where the aim is to estimate the density and abundance of a biological population. The survey methodologies covered include line transects, point transects (variable circular plots), cue counts and trapping webs.

The aim of this documentation is to tell you how to use the program, given that you already understand the concepts. Conventional distance sampling methods are described in detail by Buckland et al. (1993, 2001), and advanced methods are described by Buckland et al. (2004). These books are essential companions to the software. (See Distance Sampling Reference Books).

This documentation is available in two formats:

- a standard help version, in Microsoft HTML-Help format. This is what you see when you press the **F1** button, or choose **Help** | **Contents and Index** from the main Distance menu.
- a print-ready version, in Adobe Acrobat (.pdf) format. You can view this version by choosing **Help** | **Online Manuals** | **Users Guide** from the main Distance menu, or from the Windows Start Menu by choosing **Programs** | **Distance** | **Users' Guide**.

For the online version to display properly, you need HTML-Help software installed on your computer, and for the print-ready version Adobe Acrobat must be installed. See the System Requirements part of the ReadMe.txt file for more details.

The documentation is divided into two main parts: Users Guide and Program Reference.

The Users Guide is designed to be read (or at least scanned!) page by page. It is divided into chapters, each of which describes a different aspect of the program. It is probably easiest to read the Users Guide in Adobe Acrobat format, because some of the sections are quite long.

The Program Reference can be referred to whenever you need to know how to use a specific part of the interface. It is probably easiest to access by pressing F1

from within Distance. This automatically takes you to a page of information about the window you are currently viewing. The Program Reference is currently included as an appendix of the Users Guide.

There are also two other appendices. The first gives an overview of the internal workings of the Distance software, and contains some reference material for users wishing to tinker "under the hood" of the program. The second is the command language reference for the CDS and MCDS analysis engines.

In conjunction with the documentation, you may like to try out the sample projects. These are located in the "Sample Projects" folder, below the Distance program folder (usually "C:\Program Files\Distance 5"). These projects are referred to at various points throughout the text, and they are used as the basis for the tutorials in Chapter 3. An overview of the sample projects is also given at the end of Chapter 3.

Conventions in This Documentation

Bold text indicates a reference to an element of the user interface of Distance (for example the **Project Browser**, or **New Analysis** button), and is also used occasionally for **emphasis**.

Indicates a paragraph in the text that makes an important point relevant to the subject being discussed.

Indicates a paragraph that contains advice about how to do something in Distance. Often the advice highlights a shortcut of some kind, or explains how to do something that is quite complicated.

Indicates a paragraph that contains some incidental information.

Indicates a topic that contains some advanced material. These topics can be skipped on first reading. You should come back to them when you are familiar with the basics of Distance.

Placed above or alongside text that contains an important warning. You should definitely read this text!

Occurs before text that describes a feature of Distance that was not present in previous versions. We hope it helps users of the old software familiarize themselves with the new version.

When describing keyboard actions, **CTRL** is short for the control key. For example **CTRL-X** means hold the control and x keys at the same time.

When describing menu selection the symbol | means "and then select" – for example **Help** | **About Distance…** means select the **Help** menu and then select **About Distance…**.

Where to Go Next

We recommend that everyone start by following the guided tour in Chapter 3. New users should then read the Users Guide Chapters 4-7 – if you do this you will find using the program much less confusing! People familiar with Distance 3.5 or later should check out the section New in Distance (in Chapter 2) and should at least glance at Chapters 4-7, as required.

Advanced users will want to check out Chapters 6, 9, 10 and the appendices.

If you have yet to install Distance, you should read the release notes in the file ReadMe.rtf that accompanies the Distance setup program.

Before you start using Distance, please make sure you have read and agreed to the Use Agreement. If you wish to cite Distance in your next Nature paper, please use the Citation for Distance.

Citation for Distance

The suggested citation for both the manual and software is:

Thomas, L., Laake, J.L., Rexstad, E., Strindberg, S., Marques, F.F.C., Buckland, S.T., Borchers, D.L., Anderson, D.R., Burnham, K.P., Burt, M.L., Hedley, S.L., Pollard, J.H., Bishop, J.R.B. and Marques, T.A. 2006. Distance 6.0. Release "x"[1]. Research Unit for Wildlife Population Assessment, University of St. Andrews, UK. http://www.ruwpa.st-and.ac.uk/distance/

[1]Instead of "x", substitute the release number of the software you are using – e.g., "Distance 6.0 Release 1". You can find the release number by selecting **Help | About Distance...** from the main menu. It is important to include the release number in the citation, because results may not be absolutely identical between releases.

Distance Sampling Reference Books

The standard reference for conventional distance sampling methods is:

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. 2001. *Introduction to Distance Sampling.* Oxford University Press, London.

The book describes the methods in detail, as well as covering aspects of survey design and giving several worked examples. It updates the previous standard work:

Buckland, S.T., Anderson, D.R., Burnham, K.P. and Laake, J.L. 1993. *Distance Sampling: Estimating Abundance of Biological Populations.* Chapman and Hall, London, reprinted 1999 by RUWPA, University of St. Andrews, Scotland.

The 1993 book is still available for download over the internet at no charge, at:

http://www.ruwpa.st-and.ac.uk/distance.book/

The advanced methods in this software are described in the following book:

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. (editors) 2004. *Advanced Distance Sampling.* Oxford University Press, London.

This book describes automated survey design methods, multiple covariate distance sampling and mark-recapture distance sampling. In addition methods are described that are not currently in Distance but which we hope to include in the future, such as spatial modelling of density and adaptive distance sampling.

Staying in Touch

Distance-sampling Email List

The purpose of this list is to promote the sharing of ideas and information among researchers and practitioners interested in distance sampling techniques. It is a relatively low-volume list, and has been running since 1998.

Suitable topics for posting include:

- questions about survey design and analysis
- new methodological developments

- use of software tools (program Distance and other software)
- news about upcoming meetings, workshops and conferences where distance methods will be discussed
- jobs in distance sampling-related fields

To join this distance-sampling list, send an email to jiscmail@jiscmail.ac.uk with the following in the body of the message (not in the subject line):

join distance-sampling yourfirstname yourlastname

Replace the text "yourfirstname" with your first name and the text "yourlastname" with your last name (e.g., join distance-sampling Joan Smith)

In response, you will receive a message back that explains how to use the listserver. More information about the listserver, and an archive of messages sent to the list, are available at the list's home page

http://www.jiscmail.ac.uk/lists/distance-sampling.html

Please check the archive of previous messages before posting! Assuming you have an internet connection, you can access the archives directly from within Distance by choosing **Help** | **Distance on the Web** | **distance-sampling List Archive** from the main menu.

Program Distance Web Site

Program updates, etc. will be posted to the program Distance web site. The web address is

http://www.ruwpa.st-and.ac.uk/distance/

Assuming that you have an internet connection, you can access the web site directly from within Distance by choosing **Help** | **Distance on the Web** | **Distance Home Page** from the main menu.

Sending Suggestions and Reporting Problems

If you are having problems with Distance, please check the chapter on Troubleshooting, and also the release notes file ReadMe.rtf (**Help** | **Release Notes**). An up-to-date list of Known Problems is in the Updates section of the Program Distance Web Site. In addition, you should check the archives of the distance-sampling email list.

Once you have exhausted these possibilities, please send a message to the program authors at distance@mcs.st-and.ac.uk. If a project file is required to reproduce the problem, please export it to a zip file (**File | Export Project…**) and send it with your email. Please remember that Distance is free, so technical support is given on a "best effort" basis (i.e., we'll do the best we can, given our other commitments).

Chapter 2 - About Distance

What is Distance?

Distance is a Windows-based computer package that allows you to design and analyze distance sampling surveys of wildlife populations (for more about distance sampling, see the distance sampling reference books).

Automated Survey Design

Using Distance, you can enter or import information about your study area into the built-in GIS. You can then try out different types of survey design to see which might be most feasible. Distance can look at overall properties of the design such as probability of coverage. It can also generate survey plans from the design. For more details, see Chapter 6 - Survey Design in Distance.

Data Analysis

Once you have collected your survey data, Distance can be used to analyze it. Analysis is done in Distance using *analysis engines*, of which there are currently three: the conventional distance sampling (CDS) engine, the multiple covariate distance sampling (MCDS) engine and the mark-recapture distance sampling (MRDS) engine. For more details, see Chapters 7-10.

Use Agreement

When you install Distance, you agree to abide by the Use Agreement. A copy of this agreement is in the Distance program directory, in the file UseAgreement.txt. This can be accessed from within Distance by selecting **Help** | **About Distance…**, and clicking on the **Use Agreement** tab.

Sponsors

Distance is currently free to all users. Nevertheless, it is not free to develop and maintain! If you use Distance on a regular basis, please consider sponsoring the software. You could either make a donation towards program development and maintenance or you could finance a specific new feature that you would find of use. More details can be obtained by contacting the program authors (see Sending suggestions and reporting problems).

A list of sponsors of this release of Distance can be seen by selecting **Help** | **About Distance…**, and clicking on the **Sponsors** tab. See also the Acknowledgements section, below.

Distance Development Team

Development project coordinator

Len Thomas is an academic fellow with the Centre for Research into Ecological and Environmental Modelling (CREEM) at the University of St Andrews, Scotland.

Programmers

Len Thomas (see above)

Jeffrey L. Laake is a statistician at the National Marine Mammal Laboratory in Seattle, USA.

Samantha Strindberg is a quantitative ecologist working in the Living Landscapes Program at the Wildlife Conservation Society, New York, USA. She was formerly a graduate student in the School of Mathematics and Statistics, University of St. Andrews, Scotland.

Eric Rexstad is a research fellow with Centre for Research into Ecological and Environmental Modelling (CREEM) at the University of St Andrews, Scotland.

Fernanda F. C. Marques is a wildlife and wildland monitoring specialist with the Amazon-Andes Conservation Program of the Wildlife Conservation Society, based in Rio de Genero, Brazil. She was formerly a graduate student in the School of Mathematics and Statistics at the University of St. Andrews.

M. Louise Burt is a research fellow with Centre for Research into Ecological and Environmental Modelling (CREEM) at the University of St Andrews, Scotland.

Jon R. B. Bishop is a graduate student in the School of Mathematics and Statistics at the University of St. Andrews.

Principal investigators

Stephen T. Buckland holds the Chair in Statistics in the School of Mathematics and Statistics at the University of St. Andrews.

David L. Borchers is the head of the Research Unit for Wildlife Population Assessment, a research unit within CREEM at the University of St. Andrews.

Jeffrey L. Laake (see above)

David R. Anderson is Professor of Fishery and Wildlife Biology at Colorado State University, USA, and was formerly Leader of the Colorado Cooperative Fist and Wildlife Research Unit before retiring in May 2003.

Kenneth P. Burnham is an Assistant Leader of the Colorado Cooperative Fish and Wildlife Research Unit and Professor of Biological Statistics at Colorado State University, Fort Collins, USA.

Development team members

Sharon L. Hedley is a consultant statistician based in Fife, Scotland.

John H. Pollard is a software engineer and former graduate student in the School of Mathematics and Statistics at the University of St. Andrews.

Tiago A. Marques is a graduate student in the School of Mathematics and Statistics at the University of St. Andrews.

Acknowledgements

We are grateful to our respective agencies and institutions for their support during the production of this and the previous versions of Distance. In particular, the Biotechnology & Biological Sciences Research Council (BBSRC) and Engineering & Physical Sciences Research Council (EPSRC) provided full

funding to Len Thomas to coordinate the development of Distance 4. Additional financial support for the production of Distance 4 and 5 came from the Wildlife Conservation Society, the US National Marine Fisheries Service, the US National Parks Service, Department of Fisheries and Oceans Canada and the Colorado Division of Wildlife, and for Distance 6 came from Geo-Marine Inc.

We thank Brad Martinez, of the Common Controls Replacement Project, and Steve McMahon, author of vbAccelerator.com, for common dialog code and components. Thanks also to David Liske of Delmark Computing Services for his open source HTML help class, Francesco Balena, editor of Visual Basic Programmers Journal, for the extended collection class, Phil Fresle of Frez Systems for the compression wrapper class, and Karl E. Peterson, VB MVP, for various published code snippets.

Julian Derry, from the University of Edinburgh, helped with the programming of the first windows-based version of Distance. Katy Clarke helped to improve the manual and provide additional sample projects. Greg Fulling, Carter Watterson and Anurag Kumar, all of Geo-Marine Inc., helped test the DSM engine.

History of Distance

Distance evolved from program TRANSECT (Burnham et al. 1980). However, Distance is quite different from its predecessor as a result of changes in analysis methods and expanded capabilities. The name Distance was chosen because it can be used to analyze several forms of distance sampling data: line transect, point transect (variable circular plot) and cue-counts. By contrast TRANSECT was designed only to analyze line transect data.

Distance versions 1.0 - 2.2 were DOS-based applications that were programmed using a relatively simple command language. Version 3.0 was a windows console application, but retained the command language structure. All of these versions were principally programmed by Jeff Laake of the National Marine Mammal Laboratory, US Fisheries Service.

In 1997, Steve Buckland and David Borchers, from the University of St Andrews, obtained funding from two British research councils to proceed with an ambitious three-year project to develop new distance sampling software. The new software, which became known as Distance 4, was designed to be fully windows-based, and be capable of incorporating new features such as geographic survey design, multiple covariate distance sampling models, spatial estimation of abundance, and dual observer mark-recapture line transect methods. A Distance 4 project development team was assembled, coordinated by Len Thomas. In autumn 1997, it was decided to produce an intermediate version of Distance: fully windows based, but with the same analysis capabilities as the current version 3.0. This new program, Distance 3.5, took one full year to develop, and was released in November 1998, with various updates through to February 1999. Distance 3.5 was downloaded by over 4000 users, from around 120 countries.

Extension of Distance 3.5 to become Distance 4 began in 1999, and the software was first previewed at training workshops in summer 2000. After various public beta versions, Distance 4.0 was released in 2002, followed by Distance 4.1 in 2003 and Distance 5.0 in 2006. This last version has a major new feature in the form of a link to the free statistical software R, thereby facilitating a major expansion in the analytical capabilities potentially available to Distance users. We have also released a beta version of Distance 6.0, which contains a new density surface modeling engine

We are still actively developing the software, incorporating new features and extending current ones. If you have any comments or suggestions about the program, we'd love to hear from you!

New in Distance

New Features of Distance 6.0

New Analysis Capabilities

- New analysis engine: density surface modeling (DSM).

New Features of Distance 5.0

New Analysis Capabilities

- Interface with the free statistics software R
- New analysis engine: Mark-recapture distance sampling (MRDS). Allows analysis of dual observer distance sampling surveys, where probability of detection on the trackline can be estimated. Currently restricted to line transects. See Chapter 10 - Mark Recapture Distance Sampling in the Users Guide for details.

Other Imrovements

- Better bootstrapping in the CDS and MCDS engine: progress is given on the main toolbar; bootstrap point estimates are given in the output and analysis browser columns (useful for model averaging – see Model averaging in CDS Analysis).
- Better documentation (e.g., MCDS engine command reference in an appendix) and many more sample projects.

New Features of Distance 4.1

New Analysis Capabilities

- Better goodness-of-fit testing – q-q plots, K-S tests and Cramer-von Mises statistics for exact data in both CDS and MCDS engines
- When calculating adjustment terms, you can now scale distance by either truncation distance w or estimated scale parameter σ. More details are in Chapter 9, under Defining MCDS models (since the option mostly applies to the multiple covariate engine).
- When combining stratum estimates to form a global density estimate, if you weight by effort you can now treat the effort strata as either fixed or random effects (see Stratification and Post-stratification, Chapter 8 - Conventional Distance Sampling Analysis).

Graphical User Interface

- GIS has improved capabilities for coping with projections: the list of candidate projections is much longer and you can now specify projection parameters.

New Features of Distance 4.0

Survey Design

- Ability to enter study area details into built-in GIS

- Can then examine properties of user-specified designs, from a wide range of different design classes (e.g., random points, grid of points, line segments, zig-zag lines, etc)
- Can generate survey plans for the different designs.

New Analysis Capabilities

- New MCDS engine allows multiple covariates in the detection function.
- Data selection function of the Data Filter now much more powerful.
- For surveys with nested sample layers (e.g., clusters of points along a line), the user can choose which level to treat as the independent sample for determining encounter rate variance.
- Analysis engines now calculate AICc, as well as other model selection measures. More of these statistics are summed across strata for stratified analyses.
- For stratified analyses, MCDS engine allows estimation of probability of detection by stratum when detection function estimated globally – can reduce bias when there isn't enough data to estimate detection function by stratum.
- For stratified analyses, starting values and bounds on parameters are now specified independently for each stratum.

Graphical User Interface

- Built-in GIS, based on MapObjects by ESRI (compatible with ArcView)
- Data Explorer and Analysis Browser consolidated into the Project Browser. New tabs for Designs, Surveys and Maps added.
- New Analysis Components window makes it easier to keep track of Data Filter and Model Definitions.
- Lots of small refinements, e.g.:
  - Data Filter and Model Definition properties remember which tab they were last on, making setting up analyses quicker.
  - Naming objects is now easier – double click on the names in the Project Browser, or in the Details windows to edit them.

Distance Projects

- Can now export projects to zip archive files, and open projects directly from the archive file. Zip files are automatically spanned across removable disks (e.g., floppy disks) if they are too big to fit on one disk.
- Can set up projects using another project as a template

Data Storage

- Much more flexible data structure – data can be linked from external databases in a variety of formats (although there is no user interface for this yet – has to be done by direct editing of the database).
- Internal data now stored in DistData.mdb file within separate data folder. The data folder is also used by default to store the GIS information (ESRI shapefiles). Distance 4 project is therefore a project file (.dst file) and the associated project folder.

- More layer types, including coverage probability, nested strata and effort layers, etc. Can now have unlimited number of data layers in the project
- New survey object allows multiple surveys to be stored together in the same project file, but analyzed separately. Surveys can have different survey methods and different data layers.
- Units for data and analysis can be changed after the project is set up.

New Utilities

- Data import now more flexible – can import one or more data layers, and can use ID or Label fields to link to the parent layer. Various limitations in previous version caused by MS database engine not correctly recognizing the data types of text fields have been worked around.

New Features of Distance 3.5

Graphical User Interface

- Well defined menu structure and button-bars allow user to navigate through program
- Interactive "Wizards" guide the user through the process of setting up a new distance project
- Spreadsheet-based Data Explorer for entering data
- Summary table (Analysis Browser) allows the user to view and compare analyses, and is the starting point for creating and running new analyses
- Analyses can be grouped into sets for convenient archiving
- Analysis specification is completely graphical - users do not need to learn a command language to use the program
- Each analysis is split into two components: Data Filter and Model Definition, allowing for easy reuse of the components to create new analyses
- Results of multiple analyses can be compared side-by-side in Analysis Details windows
- Any error and warning messages generated during the analysis are clearly displayed
- Detailed results output is split into pages for ease of viewing
- Fitted detection functions are displayed as high resolution plots
- Extensive windows-based help; context sensitive help available at any point in the program

Robust Data Storage

- Data and analyses stored in single file (a distance project file), which has a robust, industry standard database structure

New Utilities

- Import of data from text files - "flat file" format allows easy export from common database and spreadsheet applications
- Import of command files from previous versions of distance

- Click one button to copy high-resolution plots, results text, analysis tables or data from Distance to the windows clipboard. From there paste into any word processor or spreadsheet.

New Analysis Capabilities

- Additional information can be stored in the Data Explorer, and this can be used to subset or post-stratify the data during later analysis
- Data Filter allows selection of subsets of the data for analysis
- Data can be post-stratified (for example by sex or species) for estimation of components of the analysis
- Multiple analyses can be run at one time (e.g., bootstrap analyses can be run in the background)
- Multipliers offer a flexible way to scale the density estimate to account for indirect counts, $g(0) < 1$, etc.
- Analysis Engine now fully 32-bit, making it significantly faster and allowing analysis of larger datasets
- Numerous small improvements and fixes have been made in the Analysis Engine

Changes in Distance Analysis Engine

- This section lists changes in the analysis engine's capabilities between Distance 2.2 and 3.5 (although we don't list all the minor fixes). There will undoubtedly be some differences in results between the versions, due to our switch from the old 16-bit NAG fitting routines to the new 32-bit IMSL routines. The switch was made for performance (speed and memory) and licensing reasons, and we expect the new routines to perform (i.e. Converge) as well, if not better, than the old ones.
- AIC is summed over strata for stratified analyses
- AIC is now the default for automatic selection of adjustment terms. This should make little difference in practice, but was done for consistency with the use of AIC to select among multiple models.
- Can set upper and lower bounds on key function parameter estimates
- Can select among multiple models using BIC
- Estimation algorithm tweaking options (EPSILON and ITERATIONS) are now gone. (The new fitting routines by IMSL don't have these options.)
- Parametric bootstrap for F0 (old VARF command) gone. (It was not very useful.)
- Cluster size estimate is now based on regression of $g(x)$ vs $\log(x)$ by default (used to be test first and use regression only if statistically significant). Nothing is lost by using the regression by default, and the method is now more consistent among analyses. We also don't like hypothesis tests much and would rather avoid them.
- The user can no longer use different adjustment term selection methods in different models within the same estimate (ie ESTIMATOR /SELECT command has gone).
- Only one ESTIMATOR /CRITERION switch is allowed in an ESTIMATE section

- The default number of bootstraps is now 999 (used to be 1000 - should make almost no difference in practice). (For justification, see Buckland, S.T. 1984. Monte Carlo confidence intervals. Biometrics 40:811-817)

Other Minor Changes

- The plot file format is now slightly different (see Model Definition Diagnostics)
- The stats file has some additional statistics (see Model Definition Diagnostics)

New Features Planned for Future Versions

Survey Design

- More design classes
- More design statistics – index of spatial spread of sample units
- Adaptive survey design

Analysis

- Analysis of adaptive survey designs
- Minor improvements:
  - Some refinements in the Model Definition properties to make the exploratory stage of analysis easier to do

Simulation

- Ability to simulate animal populations, lay surveys down upon these simulated populations and do analyses. Can then look at relative efficiency of different survey designs, evaluate the bias of estimators, etc.

Graphical User Interface

- Exploratory data analysis engine
- Easy-to-use interface for linking GIS data, and attribute data without importing it.
- Import data from non-text data sources
- Ability to manipulate display properties of maps

and more ...

Chapter 3 - Getting Started

Objective

The aim of this chapter is to provide a gentle introduction to Distance, and an overview of its capabilities. We don't go into much detail, but focus instead on giving you an impression of how the software works and where to find things. Please note that **this is not a substitute for reading the rest of the manual**. You will need to know about Distance projects (Chapter 4) and how data is stored in Distance (Chapter 5) before you can use the program effectively for either survey design (Chapter 6) or analysis (Chapter 7).

This chapter gives step-by-step instructions to walk you through four examples. In the first, the goal is to perform a preliminary analysis of some straightforward line transect data. We create a new Distance project, import the data, and do some preliminary analyses. In the second, we deal with import of slightly more complex survey data. In the third, we look at creating a geographic project and using it for survey design, and in the forth we look at more complex geographic data using one of the sample projects.

Note that this Users Guide is available in both on-line and print-ready formats. If you're currently reading the on-line version, you may find it easier to follow this chapter in the other format – see the Welcome topic for more about the different formats available.

To start with, we'll assume that you've downloaded and installed Distance. If not, go to the Program Distance Web Site for instructions.

Example 1: Using Distance to Analyze Simple Data

In this section, we will create a new Distance project, import some simple line transect data and perform some initial analyses. The data are the simulated line transect data used as a running example in Chapter 4 of Introduction to Distance Sampling (see Distance Sampling Reference Books). These data are also available as the sample project "Line Transect Example" (see Sample Projects).

Example 1 - Preparing the data for import

In general, we do not recommend that survey data be entered directly into Distance. Instead, it should be stored in some purpose-written data management software, such as a spreadsheet or database, and then imported into Distance.

The data for this example are stored in a Microsoft Excel file. Distance cannot import directly from Excel (or any other package); you have to first export the data to a text file. (If you do not have a copy of Excel on your computer, see the end of this topic for instructions.)

- Use Excel to open the file Example1.xls. This file is located in the Sample Projects folder, which is below the Distance program folder (usually "C:\Program Files\Distance 6\Sample projects").
- Note the layout of the file:
  - There is one column for each of: stratum name, stratum area, transect name, transect length, distance (perpendicular distance in this case) and cluster size. (The exact columns required depends on the survey – e.g., cluster size would not be required if objects were individual animals.)
  - There is one row for each observation. Look at row 102 -- the distance and cluster size entries are missing. This represents a transect ("Line 11") where there were no observations.
- In Excel, choose **File** | **Save As…** Under **Save as type:** choose Text (Tab delimited) (*.txt). Click **Save**, and now close Excel.
- You should now have a text file Example1.txt in the same folder as Example1.xls. You can open the text file in a text editor (e.g., Notepad) to examine it if you like.

If you do not have a copy of Excel on your computer, you can copy the file Example1Backup.txt to Example1.txt and continue.

Example 1 - Creating the Distance project

In Distance, survey data and analyses are stored in a **project**. Before you can import the data, you first need to create a new project.

- From the Windows Start Menu, choose **Programs** (or **All Programs**), then **Distance**, and click on **Distance 6.0**.
- In Distance, choose **File** | **New Project**. A window opens asking for the name of the project to create. Under **File name**, type "Example1", and click on **Create**.
- The **New Project Setup Wizard** now starts. This is designed to guide you through creating a new project. It will ask you what you want to do and give a list of options. The first option, which is already selected is to analyze a survey that has been completed. This is what you want to do, so click on **Next** at the bottom of the window.
- The next window (**Step 2**) confirms the selection you have made and gives some introductory information. Once you have read it click on **Next**.
- The next window (**Step 3**) asks about the Survey Methods. Firstly it asks what type of survey has been carried out. Make sure **Line transect** is selected. The observer configuration should be **Single observer** (double observer methods are for estimating g(0) and are covered in the Advanced Distance Sampling book). Measurement type for this example is **Perpendicular distance**, and the observations are **Clusters of objects**. When the correct options are selected, click on **Next**.
- **Step 4** asks about the measurement units that were used. In this example, distances were measured in **Metres**, the transect length was measured in **Kilometres** and the area was measured in **Square kilometres**. Choose the appropriate options and then click **Next**.

- **Step 5** asks if you wish to add in any multipliers. No multipliers are required for this data set, so click **Next**.
- Distance is now ready to set up the project, and in the last window (**Step 6**) asks you where to go next. We want to import some data, so select the option to **Proceed to Data Import Wizard**, and click **Finish**. The project is now created and the Import Data Wizard is started.

Example 1 - Importing the data

Before you import any of your own data, you need to understand how data is stored in Distance, by reading Chapter 5 of this Users Guide. But for now, you'll get an idea of how import works by following the steps below.

- If you have followed the previous steps, the **Import Data Wizard** is now open at the Introduction page (**Step 1**). Once you have read the text there, click on **Next**.
- A window will open prompting you for the file to import. Select **Example1.txt** by clicking on it and then click on **OK**.
- The Data Destination page (**Step 3**) of the wizard will open. This page asks you about where the data from your text file should be placed in the Distance database. Before you import your own data, you'll need to find out more about these options (e.g., by pressing **F1**), but for now the default options here are appropriate for this example, so click **Next**.
- The Data File Format page (**Step 4**) is now opened. In our case, the first row contains the column labels, so click on the option **Do not import first row**, then click **Next**.
- The Data File Structure page (**Step 5**) will now open up. Here you have several choice as to how you wish to proceed.
    - If, like in this example your data is set up in columns in the same order as you wish them to appear in the Distance data sheet checking the option **Columns are in the same order as they will appear in the data sheet** means that Distance will automatically assign a layer name, field name and field type to each row. Care should be taken however to check that Distance has done this correctly, as if additional columns are included in the data Distance may mis-label columns.
    - Another other option is to manually assign the names and type. This is a procedure that is worth knowing, as you will need to use it when the automatic assignment is incorrect or your data is not in the correct order to use that option. By default, the **Layer name:** row above each of the columns shows **[Ignore]**. Double-click on the **[Ignore]** corresponding to the first column (above the stratum name column). A drop down menu will appear. Select **Region**. Now double-click on the box below, which is in the row **Field name**. Again a drop down menu appears. Select **Label**. Now double-click on the box underneath in the field type row, you will see that the word **Label** automatically appears there. Next go to the **Layer name** for the Area column. Select **Region**, then select **Area**. Note that here area will be the only option to choose from as the only other option for region has already been used. Now click on the **Field type** box, and this will be automatically filled with the word **Decimal**. Continue doing

this process until each column has been assigned a field name, field type and layer name. The boxes should look like this:

Layer name:	Region	Region	Line transect	Line transect	Observation	Observation
Field name:	Label	Area	Label	Line length	Perp distan...	Cluster size
Field type:	Label	Decimal	Label	Decimal	Decimal	Decimal

Completed data file structure for Example 1

- Now click on **Next**, and then click **Finish** to import the data.

Example 1 - Studying the data in Distance

Now the data has hopefully imported successfully, and the project you created should be open with the **Data** tab of the **Project Browser** selected. This is the main interface to your data in Distance, and is called the **Data Explorer**.

- In the left-hand pane, under **Data layers**, Click on **Observation**. This expands the view in the right-hand pane to show you all the data. Scroll down to verify that all 12 transect lines and 105 observations have been imported. Note that line 11 has no observations associated with it, as we wanted.

Example 1 – Running the first analysis

Now you've created a new project and imported your data, it's time to do some analyses.

Before you run any of your own analyses, you should read Chapter 7 of this Users Guide, which contains more information about how analyses work in Distance, and lots of pictures of the various parts of the interface. You should also look at the subsequent chapters which give details of each analysis engine.

For now, however, you'll get an idea of how analysis works by following gthe steps below.

- Click on the **Analyses** tab of the **Project Browser**. This brings up a table called the **Analysis Browser**. This is where the analyses that are carried out will be listed. In the left-hand pane is a line highlighted that says **New Analysis**. This is the default analysis that Distance creates. A grey ball (the **status ball**) also on that line indicates that the analysis has yet to be run.
- Double click on the grey ball (or choose **Analyses** | **Analysis Details**). A new window titled **Analysis 1** will open up. This window is the Analysis Details window, and you are on the **Inputs** page (see tabs along the side).
- At the bottom of the Inputs page is a section called **Model Definition**. The selected model definition is called **Default Model Definition** and this is the only one that has been defined so far. Click on **Properties…** beside the model definition to open up the **Model Definition Properties** window. This window gives all the options available to change a model in Distance.
- Click on the **Detection Function** tab .This will show what key function and adjustment term are currently selected for use. Explore the other options if you wish, then click **OK** to exit the Model Definition Properties.
- Click on Run in the Analysis Details Inputs page. Once the analysis has run the **Results** tab will turn green indicating that there were no problems with the analysis, and you will be taken to the Results page. If there had been a problem, then you would be taken to the **Log** page of the Analysis Details window. Then the

log tab would be coloured either amber for a warning, or red for an error.

- In the **Results** tab click **Next** several times to view each page results. Once you have finished looking at the results, close the Analysis Browser window by clicking on the close button in the top right hand corner of the window, or choosing **Analysis - Results** | **Close**.
- In the **Analyses Browser**, the grey ball should now have turned green, to indicate that the corresponding analysis ran OK.. A summary the results is given in the right hand window pane. This is useful for when you wish to compare various analyses, without having to scan through all the results pages. (The columns shown can be configured by choosing **Analyses** | **Arrange columns…**.)

Example 1 - Creating a new analysis

The default analysis we ran in the last section uses a half-normal key function with cosine adjustments. Here, we'll create another analysis that uses a hazard rate key function with simple polynomial adjustments.

- Return to the **Analysis Browser** (**Analysis** tab of the **Project Browser**) and click on the **New Analysis** button on the Analysis Browser toolbar (or choose **Analyses** | **New Analysis…**).
- Before you go any further, give the analysis a sensible name that reflects what it does by clicking on the current name ("New Analysis 1") and then typing in the new name (e.g., "Untruncated hr+poly"). You might want to give a sensible name to the default analysis too (e.g., "Untruncated hn+cos"). This might not seem important when you only have two analyses, but if you don't name new analyses as you create them you'll soon find yourself loosing track of which is which!
- Double-click on the status ball for the new analysis (or make sure it's highlighted and choose **Analyses** | **Analysis Details**) to go to the **Analysis Details** window for this analysis. Because the analysis is not run, you are taken to the **Inputs** page.
- Click on **New…** in the **Model Definition** section. The Model Definition Properties window will open up. Choose the **Detection Function** tab and under **Models**, choose **Hazard rate** key and **Simple polynomial** adjustments.
- This new model definition is currently called "Default Model Definition 1", so give it a sensible name e.g., "hr+poly" by clicking on **Name:** at the bottom of the **Model Definition Properties**. Then click **OK**.
- You might also want to give a sensible name to the other model definition (currently "Default Model Definition"). One way to do this is to select that model definition in the Inputs page, open the Model Definition Properties, and change the name there (e.g., to "hn+cos"). Another, easier, way is to double-click the name on the Inputs page and change it there.
- Now run the analysis and compare the results with the previous one.

Example 1 - Further investigations

As we mentioned in the introduction, these data form the running example in Chapter 4 of Introduction to Distance Sampling. In the book the various analysis options are explored including truncation, grouping into intervals, various key function + adjustment combinations and different methods of estimating mean cluster size in the population.

To truncate the data a new **Data Filter** needs to be created. To understand properly what Data Filters and Model definitions are, you need to read Chapter 7 - Analysis in Distance, but for now we will show you how to create two new analyses with a new Data Filter.

- In the **Analysis Browser**, click on the **New Analysis** button to create a new analysis.
- Name this analysis something like "19m trunc hn+cos"
- Double click on the status ball to open the **Analysis Details** for this analysis.
- Under Data Filter, click **New…**.
- In the **Truncation** tab, select **Discard all observations beyond** and type in 19.
- Give the new data filter a sensible name (e.g., "19m truncation"), and then click **OK**.
- Check that the highlighted Model Definition for this analysis is the "hn+cos" one. Note that two analyses are now using the same Model Definition – this analysis and the first analysis we ran, which was without truncation.

Let's not run this analysis yet – instead we'll create a second analysis with 19m truncation and hazard-rate + simple polynomical model.

- Choose **Window** | **Project Browser** or **View** | **Project Browser** (or click on the View Project Browser button) to put the Project Browser window on top.
- Create another analysis, and open the Analysis Details for this analysis. (You now have two Analysis Details windows open.)
- Name the new analysis "19m trunc hr+poly".
- Make sure the "19m truncation" Data Filter is selected for this analysis, and the "hr+poly" Model Definition.

We could run this analysis by clicking the **Run…** button, but here's a different way to run more than one analysis at once. It's useful if you want to run a number of time-consuming analyses, so you can go off and do something else while they are running.

- Choose **Window** | **Project Browser** (or click on the View Project Browser button) to put the Project Browser window on top.
- Click on the "19m trunc hn+cos" analysis and then hold down the **CTRL** key and click on the "19m trunc hr+poly" analysis. This should select both of them. Click on the Run Analysis button on the Analysis Browser toolbar or choose **Analyses** | **Run Analysis**. The two analyses will be run once after another.
- When they are finished, you can look at the results in the two Analysis Details windows. (Note: the Analysis Details windows do not have to be open to run an analysis – you can just highlight it in the Analysis Browser and click the Run Analysis button.)

Try creating more analyses that replicate the analyses in Chapter 4 of Introduction to Distance Sampling (such as grouping the data into intervals before analysis – to do this you need to create a new Data Filter and change the options in the Intervals tab).

As you create more Data Filters and Model Definitions, you may find that you want to change their order, delete them or rename several at a time. A convenient way to do this is using the **Analysis Components** window, which can be opened by choosing **View** | **Analysis Components** or clicking the button on the main toolbar. In the Analysis Components window, clicking the first button lists the Data Filters and clicking the second button lists the Model Definitions.

An example of a set of Analyses, Model Definitions and Data Filters that have already been set up is given in the Ducknest sample project. To open this, choose **File** | **Open Project…** and choose "Ducknest".

Example 2: More Complex Data Import

In this section, we will import point transect data on house wrens (*Troglodytes aedon*). The study is used an illustrative example in section 8.6 of Introduction to Distance Sampling (see Distance Sampling Reference Books), and there is a sample project that already contains the data (House Wrens.dst - see Sample Projects) if you want to skip the import part of the section.

The data were collected from 155 points, with between 14-16 points within each of 10 study blocks. Each block was of size 16-ha. The blocks were situated in riparian vegetation along 30km of South Platte River bottomland near Crook, Colarado. The data were collected by 4 observers, who each visited each point. In the analysis, we will post-stratify by observer to recreate the results from the Distance book, but another possibility would be to use observer as a covariate in a multiple covariate distance sampling analysis.

We assume the user has already worked through Example 1: Using Distance to Analyze Simple Data, so is familiar with the basics of project creation, data import and data analysis.

Example 2 - Preparing to import the data

The data are again stored in a Microsoft Excel file: Example2.xls. Use the instructions from the previous example to save this to a tab-delimited text file Example2.txt. If you don't have a copy of Excel on your computer, copy the file Example2Backup.txt to Example2.txt.

Now create a new Distance project, Example2.dst:

- In Distance, choose **File** | **New Project**. A window opens asking for the name of the project to create. Under **File name**, type "Example2", and click on **Create**.
- Click through the **New Project Setup Wizard**.
  - In **Step 3**, you need to specify that this is a **Point transect** survey, and the measurements are therefore **Radial distance**.
  - In **Step 4**, the units of measurement are **Metres**, and area is measured in **Hectares**.
  - In **Step 5**, we do want to define a multiplier. We need to tell Distance that each point was visited 4 times. One way to do this would be to have a column in the data for survey effort, and have a value of 4 in each row. This would be most useful if the survey effort (number of visits) varied between points.

In this case, there were 4 visits to each point, so we will define a multiplier called "Visits" that divides the density estimate, and later we will put a value of 4 for this multiplier. (For more on Multipliers, see Multipliers in CDS Analysis). To define the multiplier, tick **Other** and give it the field name "Visits". Choose **N** for Create Fields for SE and DF (there is no standard error or degrees of freedom associated with this multiplier).

- In **Step 6**, choose **Proceed to Data Import Wizard**, and click **Finish**.

Example 2 - Importing the data

You can pretty-much follow the instructions from the previous example to import the data. The only change is in the Data File Structure page (**Step 5**) of the Import Data Wizard. Here there are two differences: (1) you can't use the option that Columns are in the same order as they will appear in the data sheet (because the Survey Effort field is missing), and (2) there is an extra field for Observer. To get around (1) you should manually assign the field names to the columns, as outlined in Example 1. The following are instructions for creating a new field to put the Observer data in.

- First manually assign all the other columns to the appropriate fields.
- Then in the **Layer name:** entry corresponding to the Observer column, choose Observation (the observer data is an observation-level column). Click on the entry corresponding to the **Field name:** row and the text "New Field" will appear. Type in "Observer" instead, and hit **Enter** to confirm this name. Now in the entry corresponding to the **Field type:** row, choose **Integer** (you could equally well choose **Text** as the field is likely only going to be used for data selection and possibly as a factor covariate in the MCDS engine).
- Click **Next** to continue to the final Step, and then click **Finish**. The new field will be created and the data imported.
- Once the data have been imported, use the **Data Explorer** (the **Data** tab of the **Project Browser**) to check the new field had been created correctly.

Example 2 - Analysis

Before analyzing any data, we need to enter the number of visits in the multiplier field that has been created in the project.

- Click on the **Data** tab of the **Project Browser**.
- Under **Visits**, enter the number 4.

We also need to set up the multiplier correctly in the **Multipliers** tab of the first **Model Definition**.

- Choose **View** | **Analysis Components** and then **Analysis Components** | **Model Definitions**.
- Open the default model definition's properties by choosing **Analysis Components** | **Item Properties…** or by double-clicking on the ID 1 in the **Analysis Components Window**.
- Under **Multipliers**, there should be one multiplier defined, with field name Visits. However, the **Operator** is wrong – it is set to

"*" (multiply) by default, while we want the density estimate to be divided by the number of visits, so we choose "/".

- Click **OK**. Each time you define a new model definition, it is based on the one you have currently selected, so all future model definitions will automatically now have the correct operator.

If you wish, you can now set about recreating the analyses of section 8.6 of Introduction to Distance Sampling. However, note that you'll need to learn about stratification (by area), post stratification (by observer) and data selection (selecting out each observer) to recreate all the analyses in the book. These are all outlined in this Users Guide in Chapter 7 - Analysis in Distance and Chapter 8 - Conventional Distance Sampling Analysis.

Note also that you're likely to obtain slightly different results from those in the book, as the data are slightly different and the CDS analysis engine has been modified since the book was written.

You can also try using observer as a factor covariate using the multiple covariate distance sampling (MCDS) analysis engine – for more on this see Chapter 9 - Multiple Covariates Distance Sampling Analysis.

Have fun!

Example 3: Using Distance to Design a Survey

In this section, we will design an aerial line transect survey of marine mammals in an area of approximately 100km^2 adjacent to part of the Scottish coastline, called St. Andrews Bay. We will use a systematic grid of parallel lines. The small survey plane permits a total flight length of approximately 250km, excluding the flight time to and from the landing strip further down the coast. We would like to create a design that uses the maximum possible proportion of the flight length moving along transect lines (as opposed to moving between the lines), while at the same time respecting the overall constraint of 250km.

Example 3 - Preparing the data for import

Before you can start designing a survey, you need to create a new Distance project and enter the coordinates of the boundary of your study area. There are several ways to get geographic data into Distance– here we're going to import the co-ordinates from a text file. You can also type the coordinates in manually or import them from a Geographic Information System (GIS) – see Chapter 5 - Data in Distance for more on these.

You should also read the information in that chapter about coordinate systems before deciding whether to use a GIS to apply a geographic projection to your data before importing it. In this example, we have projected the data (from the OSGB 1936 geo-coordinate system using the transverse mercator projection) into meters. Distance expects data in meters by default, so this way we won't have to deal with any of the geographic coordinate system or projection facilities in Distance.

You can examine the coordinate data in the file StAndrewsBay.txt in the Sample Projects folder, which is below the Distance program folder (usually "C:\Program Files\Distance 6\Sample projects").

If you want to skip the steps involved in creating the project and importing the data, you can open the sample project StAndrewsBay.dst (choose **File** | **Open Project…** and, select StAndrewsBay and click **Open**). Then, go to the section entitled Example 3 - Creating a Survey Design.

Example 3 - Creating the Distance project

To create a new project to contain our geographic data and survey designs, follow these steps.

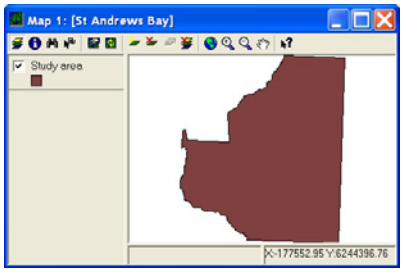
- From the Windows Start Menu, choose **Programs** (or **All Programs**), then **Distance**, and click on **Distance 6.0**.
- In Distance, choose **File** | **New Project**. A window opens asking for the name of the project to create. Under **File name**, type "Example3", and click on **Create**.
- The **New Project Setup Wizard** now starts. This is designed to guide you through creating a new project. It will ask you what you want to do and give a list of options. You want to choose the second option, to design a new survey. Then click on **Next** at the bottom of the window.
- The next screen contains some information about what Distance will do next. Read the text and click **Finish**.
- The Distance project is now created and you are taken to the Data tab of the project browser. On the left-hand pane, under **Data Layers**, you can see that one data layer has been created called "Study area". In the right-hand pane, under **Contents of Global layer 'Study Area'**, you can see that the "Study area" layer has one record, with ID 1 and Shape "Polygon".

Example 3 - Importing the Geographic Data

- Use a text file editor to open the StAndrewsBay.txt file which contains the coordinates of the study area. Highlight all of the rows of data in the file and use the text editor's copy facility to copy them to the Windows clipboard. (If you are using Windows Notepad, you would choose **Edit** | **Select All** and then **Edit** | **Copy**).
- Switch back to Distance. Double-click on the word "**Polygon**". This opens a window called the **Shape Properties** dialog.
- Chose **Paste from Clipboard**. The X and Y columns should now contain the rows you copied from the text file, and the top right hand window that says "# vertices" should now read 257 (a vertex is a corner, so this tells you how many coordinates you have pasted in).
- Click OK to close the Shape Properties Dialog.

Example 3 - Checking the Data on a Map

- Click on the **Maps** tab of the Project Browser.
- From the menu, choose **Maps** | **New Map**.
- Double-click on the Name "New Map" and rename the map "St Andrews Bay". Hit **Enter** to save the new name.
- Choose **Maps** | **View Map**. A **Map** windows opens, with title "Map 1: [St Andrews Bay].
- Choose **Map** | **Add Layer**. A window appears asking which layer to add. Since the project currently contains only one layer, click **OK** to add the Study Area to the map. You should see a window something like the following:



Map of the St Andrews Bay study area

- Click the ☒ button in the top right corner of the Map window to close the map. You will be asked whether you want to save the changes to have made to the map (you added a new layer to it) – choose **Yes**.

Example 3 - Adding a Coverage Layer

Before we can design any surveys, we need to add a second data layer to the project called a Coverage layer. This is a grid of points at which we assess coverage probability – i.e., the average probability of being covered by a survey line. For more on coverage probability and associated concepts, see Chapter 6 - Survey Design in Distance.

- Click on the **Data** tab of the Project Browser.
- Choose **Data** | **Create Data Layer**. The **Create Data Layer** dialog opens.
- Under Layer Name, type **Grid**, and under Layer type, choose **Coverage**.
- We now need to set the grid spacing. To do this, click on the **Properties…** button to open the **Grid Properties** dialog. As a rough rule of thumb when assessing coverage probability, you want a grid spacing approximately equal to the transect strip width. In our case, the truncation distance will be approximately 2km either side of the plane, giving a strip width of 4km. So, we type in 4 under **Distance between grid points** and choose Kilometre as the **Units of distance**.
- Click **OK** to close the Grid Properties dialog, and **OK** again to close the Create Data Layer dialog. Distance tells you that the grid spacing you have chosen will lead to approximately 62 grid points being created. Click **OK** to confirm.

In larger study areas, choosing a grid spacing of about the transect width causes a very large number of grid points to be created (10's of thousands). In these cases, you would want to compromise and choose a larger grid spacing.

- You can now view the coverage grid on your map. Click on the **Maps** tab of the Project Browser.
- Choose **Maps** | **View Map** to open the map of St Andrews Bay.
- Choose **Map** | **Add Layer**, choose layer name "Grid" and click **OK**. You should now see the coverage probability grid points on the map.

Example 3 - Creating a Survey Design

In this section we will create a survey design for a systematic parallel line transect survey with line spacing of 5 km.

- Click on the **Designs** tab of the Project Browser.
- Choose **Designs** | **New Design**.
- Double-click on the name "New Design" and type "5 km spacing" or some other suitable name for the design we will create.
- Now choose Designs | Design Details. The design details window for this design opens on the Inputs tab.
- Under Type of design, choose the **Line** sampler, and the **Systematic Random Sampling** class.
- Now click on **Properties…** to open the properties dialog for the systematic random line design.
- We can ignore the **General** properties for this example.
- Click on the **Effort Allocation** tab. Under **Length Units**, choose Kilometre, and under **Spacing**, choose 5. Note that Distance estimates on average 8 lines will be created and that the average on-effort length will be 226.2 km. Note also that an angle of 0 means that the lines will run west-east (90 would mean south-north).
- Note from the **Edge spacing** option at the top of the **Effort Allocation** page that we are using "minus sampling" - this means that sample lines will be placed only within the study area (plus sampling means that lines area also placed within a buffer around the outside of the study area, of width equal to the truncation distance). What effect do you think using minus sampling might have on the coverage probability at the edge of the study area? We will return to this later.
- Click on the **Sampler** tab. Choose **Units** of kilometre and **Width** of 2.
- Click on the **Coverage probability** tab. Choose **Estimate by simulation**, with 100 simulations for now.
- Click OK to close the **Design Properties** window.

Example 3 - Automated Generation of New Surveys

A design is an algorithm for laying down samplers, and a survey is a random realization of this design. In this section, we will create a random survey, based on the design we created in the previous section (i.e., with a random start point and 5km spacing between lines).

- You should be looking at the **Inputs** tab of the **Design Details** window for the design you created - the window is called "Design 1: [5km grid] Set: [Set 1]".
- Click the **Run…** button, and choose the second option, to **Create a new Survey**. Click **OK**.
- A new Survey is created, the **Survey Details** window opens, and you are taken to the **Results** tab.

- The first page "Design engine output", gives you information about this survey – for example the number of lines generated, the amount of "on effort" length (transect length) and total line length (on effort length + distance moving between the end of one line and the beginning of the next). Scroll down to look at all of the output.
- Click the **Next >** button to view a map of the samplers (lines).
- Click the **Next >** button again to get a list of the samplers' start and end points. These could easily be copied to the windows clipboard and pasted into another medium (such as a text file ready for upload to a GPS system).
- Click the button to close the survey details.

Note that the new samplers have been added to the project as a new data layer, and this new data layer has an associated GIS shapefile, so this may be a more convenient format for exporting to other systems such as GPS.

- You can verify that there is a new data layer by choosing **View** | **Project Browser**, and then clicking on the Data tab. You should see the data layer 5km grid under the **Data layers** pane.
- You can also click on the **Maps** tab, and add the new data layer to your map (or create a new map for this layer).

Example 3 - Design Statistics

So far, we have created a single realization of the design. However, since surveys are randomly generated from the design, properties such as total line length and proportion of time on effort are random properties with some unknown distribution. We are interested in knowing the average proportion of time on effort (we want to find a design where this is high), as well as the minimum total line length (we can only use designs where this is greater then 250km). To find these things out, we will ask Distance to simulate many instances from the design and record what happens each time.

- Bring the Design Details window to the foreground by choosing **Window** | **Design 1: [5km grid] Set: [Set 1]**
- Click on **Run…** again. This time, choose the option to Calculate coverage probability statistics, and click **OK**.
- We selected 100 simulations in the Design properties, so Distance does 100 simulations. The progress bar tells you what percentage have been achieved. Wait for this to reach 100%.
- The Design Details Results tab now goes green, and we can access the results.
- The first page is called **Design engine output** and contains results in text format. Note the maximum total trackline length – is it more than 250km? Note the mean on/total trackline length (this is the proportion of time on effort) – we will need this later to compare with the other designs.
- Click **Next >** to see the coverage probability map for this design. Using minus sampling, there is actually a slightly lower coverage probability within one truncation width of the north and south edges of the study area than elsewhere. However with only 100 simulations this will not be evident. Instead, you will just see random variation in coverage probability – "stripes" of high and low coverage. If you wanted to examine the edge effect in detail, you should re-run the design using many more simulations (5000,

say) and probably a finer grid spacing. If you wanted perfectly even coverage probability, you should use the "plus sampling" option for the design, in the **Design Properties**, under **Effort Allocation**.

Example 3 - Further Investigations

You can now set up more designs, with different between-transect spacings, and see if they stay within the 250km total length criterion and perhaps have a better ratio of on-effort to total line length. As a suggestion, try 4.5, 5.5 and 6km.

For more background about the survey design features of Distance, see this Users Guide, Chapter 6 - Survey Design in Distance. See also Chapter 7 of Buckland et al. (2004).

Example 4 - A Second Survey Design Project

This example provides an introduction to more of the geographic information features in Distance, using an example of survey design for 4 northwestern states in Mexico. The states form a stratum layer, so different survey effort can be allocated to the different strata. The geographic information files are not projected, so we need to specify how to project them in Distance.

The previous example (Example 3: Using Distance to Design a Survey) is simpler, and should probably be worked through first. It also demonstrates one method of geographic data import. This example is also introductory, and demonstrates different ways to access the design windows than the previous example. The two are therefore complementary.

Example 4 - Opening the Mexico Project in Distance

To start Distance:

- From the Windows Start Menu, choose **Programs** (or **All Programs**), then **Distance**, and click on **Distance 6.0**.
- Once Distance has started, on the top menu bar, select **File**, then **Open Project**.
- Select the project file "Mexico" in the dialog box, and click **Open**.

The project will take a few seconds to open. Once it's opened, you will see a window called the **Project Browser**.

The project browser is the main interface for getting things done in Distance. There are 6 tabs along the top: **Data**, **Maps**, **Designs**, **Surveys**, **Analyses** and **Simulations**. During the course of this chapter, we will examine the contents of each of these tabs, except the Simulation tab, which is disabled in this version of Distance.

Example 4 - Reviewing the Project Properties

Let's begin by examining the distance project itself. In Distance, you store all of the information about one study area in a **project**. Projects are made up of a **project file**, which is the file you clicked on to open the project, and a **data folder** (directory). We can find out more about the project in the **Project Properties** window.

- From the top menu bar, select **File**, then **Project properties**. This opens the **Project Properties** window.

The **General** tab gives you information about the location of the project file and its associated data folder (Mexico.dat).

- Click on the **Geographic** tab.

The **Geographic** tab gives you information about the default geo-coordinate system of the geographic data, and the default map projection. The geo-coordinate system is used to locate the geographic data (which is stored in decimal degrees of latitude and longitude) on the earth's surface. The projection is used to convert these data from the curved surface of the earth into a flat plane that can be used for displaying maps and designing surveys. If you are planning a survey that will take place over a small geographic area, and you are inputting your data by hand, then you don't need to worry about geo-coordinate systems or projections and can set both these options to [None]. In this example, however, the survey area is quite large and the projection chosen will make some difference to the results.

- Click on **Cancel** to close the **Project Properties** window without saving any changes you may have made.

Example 4 - Examining the Data

- Click on the **Data** tab of the **Project Browser**.

This tab contains the **Data Explorer**. In the left-hand pane, under **Data Layers**, you can see that there are four layers in the project: "Mex", "MexStrat", "Grid1" and "Grid2". You can tell the layer types by looking at the icons beside the names: Mex is a Global layer, MexStrat is a Stratum layer and Grid1 and Grid2 are Coverage layers. When you open a new distance project, the Global layer is selected by default, so the layer Mex is now selected.

- Click on the **Data Layer Properties** button on this tab to find out more about this layer.

The **Layer Properties** window opens, and under the **Geographic data** tab, you can see that the geographic data is stored in a shapefile (geographic data file) called Mex.shp in the data folder, and that the shapes in this layer are polygons (i.e. solid, multi-sided shapes).

- Click **Cancel** to return to the Data Explorer.

In the right-hand pane of the Data Explorer, you can see its' fields: ID, Label, Area and Shape. There is one record, with ID = 1. The Shape field is new in Distance 4.0 – it holds the geographic information for that record. Because this layer holds polygons, the shape record has the word "Polygon" in it.

- Double click on this word to open the **Shape Properties** window.

This is where you edit the geographic information inside Distance (an alternative is to edit the shapefile Mex.shp from outside of distance, using a GIS package such as ArcView).

- •Click **Cancel** to return to the Data Explorer.
- The coverage layers Grid1 and Grid2 contain a grid of points that will be used for determining probability of coverage for our survey designs.
- Click on "Grid 1" in the left-hand pane of the Data Explorer.

Its records open in the right-hand pane, and you can see that it has 177 records. A better way to look at the grid points is to view them in a map.

- Click on the **Maps** tab in the **Project Browser**.
- Click on the **New Map** button (3rd button along) to create a new map.
- Double-click on the words “New Map” to edit the name of the map, and call it “Coverage Grids”.
- To view the map, click the **View Map** button , or double click on the map's ID. A **Map Window** opens.

The map starts life blank – you need to add some layers to it.

- Click the **Add Layer to Map** button and select “Mex” from the list of layers.
- Click the **Add Layer to Map** button again and select “Grid 1” from the list. Now you can see the grid points.

You can also add the points from Grid 2 to the same map.

- Click the **Add Layer to Map** button again and select “Grid 2” from the list. Now you can see the grid points.

You can see that the grid points for Grid 2 are much closer together than those for Grid 1. (Grid 2 was generated with a spacing of 20 km, while for Grid 1 the spacing was 50km.) The points for Grid 2 obscure those from Grid 1 – to see both, you can change the order of the map layers:

- Press and hold the left mouse button on the legend “Grid 2” (left hand side of the map), and drag it to below “Grid 1”.

When you have finished looking at the map, you can close it:

- Click the button in the top left corner of the **Map Window** to close the map. (Say “Yes” if it asks you to save changes.)

Let’s examine the MexStrat data layer.

- Click on the **Data** tab of the **Project Browser**
- In the **Data Explorer**, click on the MexStrat layer to see those data.

You can see that there are 4 strata. If you want to see where they look like, you could create a new map in the **Maps** tab and add the MexStrat layer to the map.

When you've finished exploring the data, move on to create a new design.

Example 4 - Creating a New Design

A **design** is the template from which surveys can be created. When you specify a new design, you have to choose things like what kind of **sampler** to use (points or lines), how to place them in the survey area (e.g., random, systematic), how much effort to allocate, and how to distribute it among strata, etc. Once these are set, you use the design to generate **surveys** – realizations of the design.

Let’s create a new design.

- Click on the **Designs** tab of the **Project Browser**.
- To create a new design, click the **New Design** button . A new record appears in the left-hand pane, called “New Design”.

- Double click on this, and edit it to call the new design “150 random points”.
- Click the **Show Details** button to open the **Design Details** window.

Look under “Type of design” to see the sampler and design class; the default sampler is “Point” and the default design class is “Simple Random Sampling”.

Click the **Properties** button to set the properties for this design.

The **Design Properties** window opens. The options you see on the design properties tabs depend on the type of design. In this example, choose the following options:

- Under **Stratum layer**, choose the stratum layer “MexStrat”.
- Under **Design coordinate system**, make sure the box **Same coordinate system as stratum** is unchecked. The projection should say “Plate Carree” and the units Kilometres.
- In the **Effort Allocation** tab, under **Edge Sampling** select the **Plus** option. Uncheck the box **Same effort for all strata**. A list of the four strata in the MexStrat layer appears. Under **Allocation by stratum**, click the **Percentage from** radio button, and enter "150" as the number of points. In this example, we will put most of our effort into the two Baja strata (perhaps because this is where we think most of the animals of interest live). Under **Effort %** enter 10 for Sinaloa, 10 for Sonora, 40 for Baja Sur (south) and 40 for Baja Norte (north).
- In the **Sampler** tab, select “Kilometre” (or “Kilometer”) for the units. Let's imagine we're surveying for a very vocal species and that our truncation distance is 5km, so we enter 5 under radius (for this example we'll assume same sampler properties for all strata).
- Lastly, in the **Coverage Probability** tab, click on **Estimate by simulation** and enter 50 as the number of simulations. This is far too few for an accurate simulation, but will do for the purposes of demonstration. Under **Grid layer:** choose “Grid 2”, which is the one with the grid points closer together.
- Now click **OK** to close the **Design Properties** window. The properties window closes and we are back with the **Design Details**.

Example 4 - Automated Generation of New Surveys

- Click the **Run** button on the **Design Details** window.

A window pops up offering you two choices: (1) Calculate coverage probability statistics, and (2) Generate a new Survey.

- Choose the second option, and give the new Survey a useful name like “150 points survey” and the new layer a name like “150 points”.
- Then click **OK**.

A **Survey Details** window opens, and the status bar at the top of the Distance window says “Running Survey”. At this point you have to be patient while the survey runs. Distance is creating a set of randomly located survey points, based on the design. When it is finished, the

Survey Details Results tab opens, and you can review some statistics about the new survey.

- Click the **Next** button to see a map of the points – you should be able to see that there are more in the Western strata (Baja) than the eastern.
- Click **Next** again to see a list of the points. You can copy this into the windows clipboard by typing **CTRL-A** (control key and A key at the same time – select all), then **CTRL-INS** (control and insert – copy to clipboard). You can then paste this list into a spreadsheet, document, etc.
- Click on ☒ to close the **Survey Details** window
- Click on the **Surveys** tab of the **Project Browser**. You can see that your new survey has been added in the table of surveys.
- Select the survey and click the **Show Details** button you get back to the **Survey Details** window's **Results** tab. You are automatically taken to the Results tab of a details window that has a green status light.
- Click on the **Inputs** tab and then **Properties** button.
- Click on the **Data Layers** tab, and you can see that the new Sample data layer "150 points" has been entered as the lowest sample layer.
- Close the **Survey Properties** and **Survey Details** windows, and click on the **Data** tab of the **Project Browser**. You can see that the new sample data layer "150 points" has been added below the "MexStrat" data layer.

There is another way to generate a new survey from a design. You can do it from inside the **Survey** tab of the **Project Browser**.

- Click on the **Survey** tab of the **Project Browser**.
- Click on the **New Survey** button .

A new survey is created in the **Survey Browser** table, with the name "150 points survey 1". The first column of the survey table shows a status light, which is grey: the survey has not been "run" yet. The second column gives the ID number of the survey. The third shows which Design the survey is related to – if you hold your mouse over that number it will give you the name of the design.

Now, you want to run the survey.

- Click the **Run selected surveys** button .

The status light turns to a running person, and after a while turns green to show that the survey has been run. You can now compare the results of this survey with the last one:

- Highlight both surveys in the **Survey Browser**. To do this click on one of them, hold the **CTRL** key and click on the other.
- Now click on the **Show Details** button . The Survey Details for both surveys open, on the Results pages.
- Resize and move the windows so that they are lined up side by side, and click the **Next** buttons so that both maps are showing.

As you can see, both surveys are *realizations* of the random survey design.

Example 4 - Design Statistics

- Go back to the **Design Details** window for your design, and click **Run** again.
- This time, choose the top option (**Calculate probability of coverage statistics**) and click **OK**.

You have to wait while Distance generates multiple simulated surveys and uses these to work out the probability that each grid point will be covered by the survey. This takes much longer than generating a single survey. When it has finished, you can see the results in the **Results** tab.

- Click **Next** to see a map of the estimated coverage probabilities.

In theory, this design should produce an even probability of coverage within stratum. However, you can see that there is considerable variation. Why is this? What would happen if you repeated the run with more simulation runs (say 500, or 5000)?

Example 4 - Further Investigations

If you want to, go ahead and try out some of the other designs. For example, try a systematic grid of points. Systematic designs produce more even distribution of samplers than simple random designs, and we usually recommend them for that reason. This, and other survey design issues are discussed in Chapter 7 of both Buckland et al. (2001) and Buckland et al. (2004).

For more background about the survey design features of Distance, see this Users Guide, Chapter 6 - Survey Design in Distance.

Sample Projects

Distance comes with a number of sample projects, listed in the table below. The projects located in the "Sample Projects" folder, below the Distance program folder (usually "C:\Program Files\Distance 6", or the equivalent in your language). They demonstrate various aspects of the program – for more information about a project, open it in Distance, and choose the menu item **File | Project Properties**.

Feel free to use the sample projects as a test bed for learning about the program – try adding and deleting data, creating and running designs and analyses. Have fun!

Project name	Description
Line transect example	Simulated line transect data from Chapter 4 of Introduction to Distance Sampling. Exact data, individuals as clusters and no stratification. Step-by-step instructions for setting up a project identical to this and importing the data are given here in Example 1: Using Distance to Analyze Simple Data.
Point transect example	Simulated point transect data from Chapter 5 of Introduction to Distance Sampling. Exact data, no clusters or stratification.
Stratify example	An example of stratified data. Two strata, one with high sample coverage and one with low. Distances are exact and objects are clusters. The example is based on cetacean data, and there is also a multiplier defined to account for g(0)<1, based on a separate experiment to estimate g(0).
Ducknest	An example of how to enter and analyze interval data. There are also 8 example analyses set up, in two sets. Data are a subset of the Monte Vista duck nest data used as in

	illustrative example in section 8.4 of Introduction to Distance Sampling. The project is also set up with a suite of analyses, data filter and model definitions to illustrate a possible approach to naming and organizing these analysis components.
House wren	Point transect data on house wrens (Troglodytes aedon) in Colorado used in section 8.6 of Introduction to Distance Sampling. The project illustrates stratification, use of additional fields (observer) and a multiplier (number of visits) – import of these data is covered here in Example 2: More Complex Data Import.
Songbird	Point transect data from a multi-species survey of songbirds in Colorado used in section 8.7 of Introduction to Distance Sampling.
Golftees	St Andrews golf tee data in the format used as a running example to illustrate double-observer data in Chapter 6 of Advanced Distance Sampling.
Amakihi	An example of multiple covariate distance sampling data and analyses. Point transect data on Hawaii Amakihi, a generalist Hawaiian honeycreeper. There are three potential covariates: observer (obs), time in minutes (MAS) and hours (HAS) after sunrise. See the Project Properties for more information. This data set is the illustration data used in Marques et al. (2007), and the project contains the analyses presented in Table 2 of that paper.
StAndrewsBay	An example geographic project that can be used for survey design, comprising geographic data for the waters just off St Andrews, Scotland. Step-by-step instructions for creating a project like this one from scratch, including importing the geographic data, and also for creating example survey designs are given in this Chapter under Example 3: Using Distance to Design a Survey.
Mexico	An example geographic project that can be used for survey design, comprising geographic data for 4 states in North-Western Mexico.
LinkingExample	An example of how to link external databases and text files to a Distance project – an advanced technique outlined in Linking to Data from Other Databases.

Some of the projects contain data used in the distance sampling text books, and you can use these to recreate analyses in the books as a learning exercise. Note, however, that you may find minor differences in results between the book and the distance projects. In some cases the data in the projects are slightly different; in others differences will be due to changes in the Distance analysis engine since the books were published.

If the sample projects become mangled beyond recognition, you can restore them to their original state by re-installing them from the Distance setup program. Run the setup program, and in the window called "Select Components", uncheck all the options *except* **Sample projects**.

Chapter 4 - Distance Projects

Introduction to Distance Projects

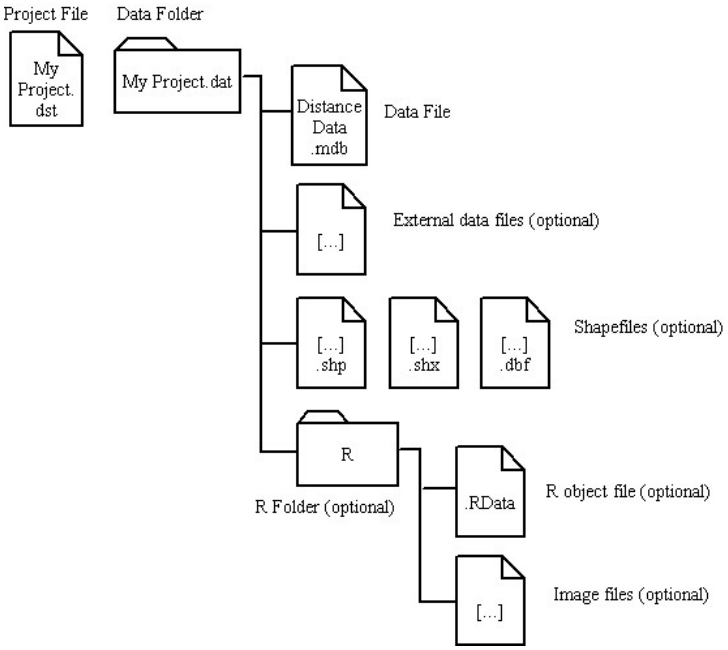
In Distance, you store all of the information about one study area in a **project**. There are two types of project: geographic projects, which contain geographic (GIS) data, and non-geographic projects, which do not. If you want to work with the survey design facilities in Distance then your project must be geographic; if you're only interested in data analysis then it can be either.

A project is made up of two parts:

- a **project file**, which contains information about the project settings, survey designs, analysis specifications and results. Project file names always end in .dst – e.g., Ducknest.dst. You can open a project file in Distance by double-clicking on it.
- a **data folder** (directory), which contains the survey data, effort data, and other related information. Data folder names always have the same beginning as the associated project file, but end in .dat – e.g., Ducknest.dat.

The data folder in turn contains one or more files, and optionally one folder:

- the **data file**, DistData.mdb. This file contains information about how the data is stored, and may contain some or all of the data itself.
- in addition, some of the data may be stored in other external data files
- if the project is geographic then the data folder will contain one or more **shapefiles**. A shapefile is a standard format for storing geographic information, invented by the GIS company ESRI. Each shapefile is actually 3 separate files: an .shp file, a .shx file and a .dbf file. (In addition, there may be other files such as .prj files.)
- if the project contains analyses that have been run using the statistical software R (e.g., those using the MRDS analysis engine), it there will be an **R folder**. This folder contains a R object file (called .Rdata), and files of images produced by R. (For more about R and the R folder, see R Statistical Software in Chapter 7).



The structure of a Distance project.

In the Windows interface, project files have the following icon:

Distance project icon

Double clicking on files with this icon in Windows opens the associated project in a Distance session. Windows also stores a list of your recently used projects in the Windows taskbar Start menu, under Documents.

You can use Windows to rename, move, copy and delete project files just like any other file – but if you do this you should do the same to the data folder. For example, if you want to copy a project from one computer to another, you should copy both the project file and the data folder.

If you're moving projects between computers, it's best to pack them up into one file first, using the export facility in Distance (see Exporting, Transporting and Archiving Projects.)

Creating a New Project

To create a new project in Distance, choose **File | New Project** on the main window menu. Alternatively, press the **New Project** button on the toolbar, or use the keyboard shortcut **CTRL-N**.

The **New Project** dialog will open, prompting you to choose a filename for saving the Distance project file. This filename also forms the basis for the title of the distance project, so it is best to choose a short, descriptive name such as "Yellow Warbler Site 1". Any characters that make an acceptable Windows filename (e.g., spaces and numbers) are allowed.

You can change the default folder that Distance uses to create and open projects by selecting this folder in a **New Project** or **Open Project**

dialog and then checking the box **Save this folder as default for Distance projects**.

After you have chosen a filename, click the **Create** button. The new project file is created and the Setup Project Wizard opens. This wizard guides you through the process of setting up the new project ready for use. Using the wizard, you can:

- set the project up ready for doing data analysis
- set the project up ready for survey design
- use an existing Distance project as a template
- import data and options from a previous version of Distance
- bypass the wizard and set up the project manually

See the Setup Project Wizard section in the Appendix - Program Reference for more about these options. Use of an existing project as a template is also described in the next section. For more about project import, see Importing from Previous Versions of Distance.

If you choose to set the project up ready for survey design or data analysis, or use an existing project as a template, then once you complete the wizard, you are ready to start entering or importing data. See Chapter 5 - Data in Distance for more information about how this is done.

Using an Existing Project as a Template

Distance can automatically set up a new project by copying the project settings, data structure, and design, survey and analysis specifications from an existing project. All that is then needed are the new data, which can be brought in by the Import Data Wizard.

Examples where this facility is useful include:

- where you want to set up multiple projects with an unusual data structure (such as extra data fields not created automatically in the Setup Project Wizard).
- where you repeatedly use a standard set of analysis specifications in your projects
- where you want to make it easy for naive users to set up projects with a standard data structure and, and give them some example analyses specifications

You can use any Distance project as a template, but you may find it easier to save the projects used as templates to a special folder, to make it easy to distinguish them from your other projects. You can save a project to another folder using **File | Export Project** - see Exporting, Transporting and Archiving Projects. When exporting projects to the templates folder, you can save space by choosing the options to exclude the data and results from the exported project.

By default, when you choose **Use an existing Distance project as template**, Distance looks in the folder "Templates" under the Distance program folder. You can change this by selecting the new folder you want to use, and checking the box **Save this folder as default for opening template files**.

Opening an Existing Project

To open an existing project, choose **File | Open Project** on the main window menu. Alternatively, press the **Open Project** button on the toolbar or use the keyboard shortcut **CTRL-O**. The **Open Project** dialog will open, prompting you to choose the project to open.

Tip! You can change the default folder that distance uses to create and open projects by selecting this folder in a **New Project** or **Open Project** dialog and then checking the box **Save this folder as default for Distance projects**.

When you open a Distance project, the Project Browser is automatically restored to the same size, position and tab as when you last closed the project.

You can also open projects that have been archived in a zip file (see Exporting, Transporting and Archiving Projects for more about archiving projects). In the **Open Project** dialog, select **Zip archive files (*.zip)** from **Files of type:**. Distance will then prompt you for the folder to extract the project files into, and will then open the extracted files.

Note! You cannot open a project if any of the files in it are read-only. If you try to do this you will get an error message. This means that you cannot open a project directly from a CD – copy it to your hard drive first.

Tip! While you're working with a project, Distance reads and writes to the project and data files constantly. It is therefore much better to keep your project on your local hard drive, where access times are much faster. We don't recommend accessing projects over a network if you can help it, and we certainly don't recommend working with projects stored on floppy disks, zip disks, etc. In all cases, you're best to copy the project to your local hard drive before opening it.

Saving and Backing Up a Project

Saving Projects

In Distance, almost all of the changes you make are "live"- that is, they are saved in the Distance Project the instant you make them. For example, if you add some new data or create a new analysis, this data or analysis is instantly recorded in the project file. Because of this, there is no need to "save" distance projects in the same way that you might save word processor files. Everything is automatically saved for you as you go along.

Even though you don't need to save your work, it is important to make backup copies. This is discussed in the next section.

Backing up Projects

Why Back Up?

Having backup copies of your Distance projects can be useful for two reasons:

1. If the Distance program quits suddenly, for example if there is a power cut or a program crash, there is a small chance that your original project may become corrupted and unusable.
2. If you make a mistake in the Data Explorer, such as deleting a column of data, there is no way to get this data back. It is instantly deleted from your project, and there is no undo facility.

In both cases if you have a backup copy of your project, such mishaps need not become disasters!

Distance has an automatic backup facility that is designed to deal with the first type of need. There is also a facility for making manual backups and restores that can be used for the second case. These are detailed below.

Automatic backups in Distance

When you open a project, Distance automatically creates a backup copy of the project file and project folder. These are created in the same directory as the original and has the same names, except that the symbol "~$" is appended. For example if your project file is "MyProject.dst", the backup project file will be "$~MyProject.dst", and the backup data folder "$~MyProject.dat". (Next time you open a project in Distance, you can use the Windows Explorer to see this backup file being created.)

When you close a project, Distance automatically deletes the backup copy of your project file. If, however, Distance exits suddenly without closing the project, the backup copy is not deleted. This means that even if your original project file is corrupted, the most work you can loose is the work you did in that session.

You can turn off this auto-backup feature in the Preferences dialog.

Automatic restore from backup

Whenever you ask Distance to open a project, it first checks for the existence of a left-over backup file. For example if you ask to open the file "MyProject.dst", Distance will check for the existence of "$~MyProject.dst". If it finds a backup, it will display the following message:

[picture of backup file message goes here]

If you choose **Yes**, Distance will open the backup (and, in the course of opening the backup will make a backup of that, called "$~$~MyProject.dst"!).

If you choose **No**, Distance will try to open the original project file. During the opening procedure, it will overwrite the old backup file. If the original project file is corrupted you will therefore have no backup. Because of this, it is not recommended that you choose **No** unless you are sure that the original file is not corrupted.

Rather than using the automatic restore feature in Distance, see Recovering From Unexpected Program Exit in Chapter 12 for a better recovery strategy.

Manual backup

Distance has no undo facility. If you accidentally delete some data or analyses, they are gone for good. Because of this, it is wise to make regular backups of your project, especially before you make any major changes.

To backup your project in its current state, choose the menu option **File | Copy project to backup**, or press the icon, or use the keyboard shortcut **CTRL-S**.

Manual restore from backup

If you made a mistake, such as accidentally deleting some data, and you have a recent backup, you may want to revert to the backup copy. To do this, choose the menu option **File | Revert to Backup Copy**.

Permanent backups

As explained above, the backups created automatically by Distance, or manually using **Copy project to backup**, are destroyed when the project is closed. Therefore it is worth occasionally making an archive backup of the project, and possibly storing this archive on another computer or semi-permanent medium such as CD. To do this, choose the menu option **File | Export Project**. See the next section Exporting, Transporting and Archiving Projects for details.

A final note on backups

At the risk of stating the obvious... Don't rely on Distance (or any other single piece of software) to keep the only copy of your precious data! If you have paper data sheets, make a photocopy and keep them at another location. If you entered your Distance data via the keyboard, use the **Copy to Clipboard** facility in Distance to copy the data into a spreadsheet or text file, and keep that file on another computer at a different location.

Exporting, Transporting and Archiving Projects

Distance allows you to export projects to another location on your computer. As well as simply copying the project file and associated data folder, Distance can export the project to a single, compressed archive (zip file). In addition, you can choose to exclude certain parts of the project from your export - for example to exclude the data, or the results.

This facility is useful in three contexts:

1. Making a permanent backup of a project. In this case, you would probably want to export the project to a zip file, to save disk space. You can then archive the zip file onto a CD, tape, disk or backup computer.
2. Transporting projects between computers. Again, you will probably want to export the project to a zip file – both to save disk space and because a single file is more convenient to move around than a distance project.
3. Making templates for setting up new projects. Here, you will want to export the project as a Distance project. For more about templates Using an Existing Project as a Template.

To export a distance project, select the menu item **File | Export Project**. For more about the options, see the Export Project Dialog section of the Program Reference.

Only the distance project file and files in the data folder are exported. If your project links to GIS files or other database files outside the data folder, these will not be included

To transport a project between non-networked computers, you can export it direct to floppy disks. As projects tend to be large, it is best to export to zip archive. If the zip archive is too large to fit on one disk, Distance will automatically span the archive across multiple disks.

Tip! When transporting projects (e.g., on floppy or over the internet), you can save even more space by excluding the results from the export as these can be re-created on the new machine – although if your results include time-consuming operations such as simulations or bootstrapping, you probably don't want to do them again!

Viewing and Editing Project Properties

You can view general information about your project in the **Project Properties Dialog**. This information includes the name, location and size of the project file, and the contents of the data folder. There is also space for you to make notes about the project. The Geographic tab allows you to turn a non-geographic project into a geographic one (but not visa-versa), and to set the default coordinate systems for the project.

To open the **Project Properties** Dialog, select the menu item **File | Project Properties**. For more information about the dialog options, see Project Properties Dialog in the Program Reference.

Compacting a Project

Advanced Topic

In a Distance project, the Project File and Data File are actually database files. Like all database files, they tend to grow as you use them. This is because records and queries that Distance no longer needs are marked as deleted, but are not actually removed from the database. Permanently removing deleted records is called "compacting" a database. To do this, the database must be closed.

Distance projects are automatically compacted when you close the project. However, if you work with an open project for a long time, you might want to compact it occasionally. To do this, choose **Tools | Compact Project**. Distance will close the project, compact it, and open it again.

Tip! Compacting a project is also a way of fixing database corruption. For more details, see Fixing a Corrupted Project in Chapter 12.

Importing from Previous Versions of Distance

Importing Distance 4 and 5 Projects

Distance 6 can import all the information contained in Distance 4.0 – 5.0 project files, including the data, designs, surveys, analyses and results.

To import a Distance 4 or 5 project, either:

- Open the Distance 4 or 5 project from inside Distance 6. A dialog box asks you whether you would like to upgrade the project, and if so whether you would like to save the upgraded project under a new name. If you click **Yes**, you are prompted for the name of the new project file. The new project is created and the information from the old project is imported. The old Distance 4 project remains unchanged in its original location. If you click **No**, the old

Distance 4 or 5 project is upgraded *in situ*, and can no longer be opened in Distance 4 or 5.

- Create a new project from inside Distance 6. In the project setup wizard, choose the option to **Import a project or command file created in a previous version of Distance**, and press **Next**. Distance then prompts you for the location of the old project file. If you select a Distance 4 project and click **Finish**, all information from that project is imported into the new project.

Importing Distance 3.5 Projects

Distance 6 can import data and project settings from Distance 3.5 project files. It will not import the data filters, model definitions or analyses.

To import a Distance 3.5 project, either:

- Open the Distance 3.5 project from inside Distance 6. A dialog box prompts you for a filename for the new Distance 6 project. When you press **OK**, the project data and settings are imported into the new project, and this new project is opened. The old Distance 3.5 project remains unchanged in its original location.
- Create a new project from inside Distance 4. In the project setup wizard, choose the option to **Import a project or command file created in a previous version of Distance**, and press **Next**. Distance then prompts you for the location of old project file. If you select a Distance 3.5 project and click **Finish**, the project data and settings are imported into the new project.

Importing Distance 2.2 - 3.0 Command Files

This version of Distance cannot import Distance 3.0 or earlier command files. This facility will be added to the full version before release. Meanwhile, to import old command files, you first have to import them into Distance 3.5, and then import the Distance 3.5 command file.

Alternatively, if the data is already in flat-file format, or is stored in a separate spreadsheet or database, you could consider setting up a new project and using the data import facilities in Distance – see Data Import in Chapter 5.

Chapter 5 - Data in Distance

Data Structure

This section describes how your survey and associated data are represented in Distance. It is *essential* reading for anyone using the program. There is quite a lot of material and jargon to absorb, but it is important to understand the concepts here in order to make efficient use of the software.

Data Layers

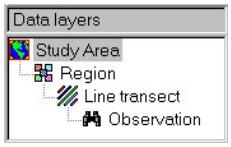
Introduction

Data in Distance are divided into a set of nested data layers. Each data layer can be thought of as a database table, with records (rows) and fields (columns). Data layers have three attributes associated with them:

- **Layer Name** (e.g., Study area, Point transect, New Layer 1) – this is a description of the layer. You can change this from the default to make it more relevant to your study.
- **Layer Type** (e.g., Global, Sample, Coverage) – this is a description of the function of the layer, and its place in the hierarchy of layers. The layer type is used internally by Distance and once it is set you cannot change it.
- **Geographic** (Yes/No). If the project is geographic, then each data layer can contain geographic information, although not all layers have to. You can tell if a layer is geographic because it will contain a Shape field (see Data Fields, below).

Hierarchy of Data Layers

Data layers are linked together in a hierarchy, with a layer of type Global at the top, and other layers below it. Here's a simple example:

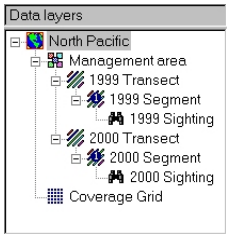


Example set of data layers from a Distance project

In this picture, the icon (symbol) tells you the layer type, while the text tells you the data layer name. The top layer, *Study Area* is of type Global. It has one child layer, *Region* of type Stratum. This in turn has one child layer, *Line transect*, of type Sample. Lastly, line transect has one child layer, *Observation* of type Observation.

Multiple Child Layers

Data layers can have more than one child layer. In the following example, the Global layer, called North Pacific, has two child layers. The first, Management Area, is of type Stratum. The second, Coverage Grid, is of type grid. The Management Area layer in turn has two child layers, 1999 Transect and 2000 Transect, both of type Sample.



More complex example data layers from a Distance project

Nested Stratum and Sample Layers

In the above example, the two Sample layers, 1999 Transect and 2000 Transect, both have layers below them of type SubSample1. This is because in this case the transect lines are divided up into segments. Each segment represents one day of shipboard surveys. By storing the effort data in separate segments, we can choose at the analysis stage whether to treat each transect as an independent sample, or each segment. For more on this, see Chapter 7 - Analysis in Distance.

In general, Sample layers can have up to 5 sub-sample layers below them: SubSample1 – SubSample5. Similarly, Stratum layers can have up to 5 sub-stratum layers below them. We hope to incorporate the ability to analyze survey designs with multiple nested stratum and effort layers into a future version of Distance.

List of Data Layer Types

A full list of layer types, together with their associated icon and the allowable child layer types is given below:

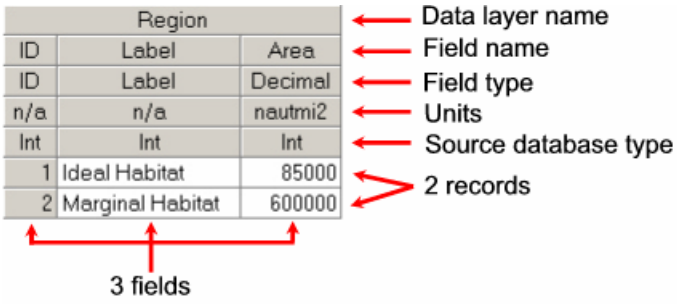
Layer type	Icon	Allowable child layer types
Global – has only one record; contains information that applies to the whole study		Stratum, Sample, Coverage, Other
Stratum – contains one record for each stratum		SubStratum1, Sample, Other
SubStratum1		SubStratum2, Sample, Other
SubStratum2		SubStratum3, Sample, Other
SubStratum3		SubStratum4, Sample, Other
SubStratum4		SubStratum5, Sample, Other
SubStratum5		Sample, Other
Sample – contains one record per sample (e.g., transect)		Observation, SubSample1, Other
SubSample1		Observation, SubSample2, Other
SubSample2		Observation, SubSample3, Other
SubSample3		Observation, SubSample4, Other
SubSample4		Observation, SubSample5, Other

SubSample5		Observation, Other
Observation – contains one record per observation		Other
Coverage – used in survey design module, for storing probability of coverage. Contains one record per coverage grid point.		Other
Other – omnibus layer type used to store miscellaneous data		Other

Data Fields

Introduction

As we said earlier, each data layer can be though of as a database table, with a number of fields (columns) and records (rows):



Example Stratum layer table

Each field has four attributes associated with it:

- **Field Name** (e.g., Area) – this is a description of the field. You can change this to almost anything you like (although some words and characters are not allowed – see Valid Field Names in the Inside Distance Appendix). (Note – you can only change the field names of internal fields – see Source Database Type, below).
- **Field Type** (e.g., Integer) – the type of data contained in records in that field. This is set when the field is created and you cannot change it. The possible field types for fields you can create are: Text, Integer, Decimal, Label. Other field types you can see are ID and Shape. More on these later.
- **Units** (e.g., Hectare) – the units of measurement of the data, where appropriate. A large number of units are available. For fields where there are no units, this should be left as n/a (not applicable).
- **Source Database Type** – tells you where the data field is located. Int (internal) means its in the project Data File; Geog (geographic) means its in a shapefile, and Ext (external) means its located in some external database.

The ID Field Type

Each layer has a field of type (and name) ID. This is used to identify each record in the layer. ID fields cannot be edited or deleted by the user.

A second function of the ID fields are to link together records in different layers. In the following example, the sighting with ID 40 in the Observation layer is

contained within the transect with ID 14 in the Line transect (Sample) layer, which in turn is contained in the region with ID 2 in the Region (Stratum) layer. The Global layer always contains just 1 record with ID 1, which encompasses the whole study site.

Contents of Observation layer 'Observation' and all fields from higher layers

Study Area		Region			Line transect			Observation			
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance	Cluster size	Clus
ID	Label	ID	Label	Decimal	ID	Label	Decimal	ID	Decimal	Integer	In
n/a	n/a	n/a	n/a	nautmi2	n/a	n/a	nautmi	n/a	nautmi	[None]	[N
Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	
1	Stratify example	1	Ideal Habitat	85000	13	13	80	33	0.31	2	
								34	0.58	2	
								35	0.49	1	
								36	0.46	2	
								37	0.36	2	
								38	0.09	2	
								39	0.03	2	
		2	Marginal Habitat	600000	14	14	75	40	0.49	1	
								41	1.94	8	
								42	1.1	10	
								43	0.85	5	
								44	0.63	7	
								45	0.39	3	
								46	0.65	1	
								47	1.16	2	
								48	0.91	2	
								49	0	2	
								50	1.18	10	
					15	15	100	51	0.1	1	
								52	0.37	1	
								53	0.59	2	
								54	0.45	2	
								55	0.53	1	
					16	16	110	56	0.21	2	
					17	17	100				
					18	18	125	57	0.85	2	
								58	1.48	2	

Example data sheet.

Note! Notice that there are no Observation records opposite the Sample (Line transect) records 16, and 17. This is because no animals were sighted on those transects.

The Label Field Type

You may also notice in the above picture that all of the data layers except the Observation layer have a data field with field type (and name) Label. This field is always created by default and can be used to name each record (e.g., "Ideal habitat" and "Marginal habitat" in the Stratum data layer). If you manually add new data layers to the project, you can add a new field of type Label if you want.

The Shape Field Type

If the data layer is geographic, then it will contain a field of type (and name) Shape. The records for this field contain a word describing the type of shape that is stored in them (Polygon, Line or Point). You can double-click on the cells in the data sheet to edit the shape in the Shape Properties Dialog.

Other Hidden Field Types

Advanced Topic

There are two field types that you won't see from the Distance data sheet: LinkID and ParentID. LinkID fields are used to join geographic and external data to that stored in the Data File DistData.mdb. ParentID fields are used, together with the ID field, to link together records from different layers. You don't normally have to worry about either of these fields, unless you are editing the distance database by hand from outside of Distance – for more on this see the Appendix - Inside Distance.

Changing the Data Structure

If all you want to do is analyze a straightforward data set, using conventional distance sampling methods, then you will probably never need to add new data layers or fields to your project file, beyond those created automatically by the Project Setup Wizard. You may want to rename the layers and fields, which you can do easily by double clicking on the names in the Data Explorer.

However, in most other cases, you will want to change the data structure, by adding new layers and fields, and sometimes deleting layers and fields. Here are some examples:

In analyzing a multi-species survey, you may want to add an additional field for Species into the Observation layer. You can then do different analyses, selecting out only the species of interest.

If you are using the Multiple Covariates Distance Sampling (MCDS) analysis engine, you will want to have extra fields for the covariables you will be adding to the detection function.

Before you can use the survey design module to examine probability of coverage, you need to add a layer of type Coverage to the project, and populate it with a grid of points.

The survey design module can be used to create new survey plans, which are stored as new data layers. After a while you may find you have too many layers lying around, and want to delete some to clean up.

All of the work of adding, renaming and deleting data layers and fields can be done from the Data Explorer. Advanced users can also change the data structure from outside of Distance, by directly editing the Data File, DistData.mdb, using a database package.

Getting Data Into Distance

There are currently five ways to get data into Distance:

- Enter data from the keyboard, using the Data Explorer.
- Enter data from the keyboard, using the Data Entry Wizard.
- Import data from a text file, using the Import Data Wizard.
- Import data from a previous version of Distance, using the Setup Project Wizard.
- Link an existing database file (including GIS shapefile) to the Distance database, by editing the Data File by hand.

While it is possible to enter data from the keyboard, it is relatively slow and inefficient for all but very small datasets. We did not have the time or resources to completely reinvent the spreadsheet! Therefore, for anything other than small datasets, you will likely want to enter and store your data in a separate spreadsheet or database application and use the Data Import facility to bring the data into Distance. This has the added advantage that your data is backed up elsewhere should things go wrong!

The following two pages cover Data Entry and Data Import in more detail. For more information about Importing data from previous versions of distance, see Importing from Previous Versions of Distance in Chapter 4. For more about linking external data, see Advanced Data Topics, below.

Data Entry

Distance provides two very similar forms for data entry via the keyboard: the Data Entry Wizard and the Data Explorer.

The Data Entry Wizard guides the user through the process of data entry. It provides on-screen advice via a text window at the top of the wizard, and moves through the data one layer at a time. It is most suitable for beginning users, but its limitation is that it can only be used on simple datasets – those with four data layers (Global, Stratum, Sample and Observation). The Data Entry Wizard opens by default at the end of the Setup Project Wizard, if you are setting up a project for data analysis.

The Data Explorer is not so structured - the user is free to jump among the data layers, and there is no panel of advice. It is more suited for checking data, and for data entry by more advanced users.

For more specific information about the layout of these forms, and how to use them, see the Project Reference pages on the Data Explorer and Data Entry Wizard.

Data Import

This section contains information about importing non-geographic data. For information about GIS data in Distance, see Geographic (GIS) data, and for information about importing this data see Importing Existing GIS data.

Introduction to Data Import

Distance can import data from text files straight into your Distance project. The data in the text file should be in "flat file format" - i.e., arranged in rows and columns.

The simplest case is where you want to import survey data into a new project, after the project has been set up for analysis by the Setup Project Wizard. In this case, the project will contain four data layers: Global, Stratum, Sample and Observation (these are the layer types, the layer names will be different – e.g., the Global layer is called "Study area" by default). At a minimum, your text file should contain the following columns:

- stratum label column (you can get away without this – see below)
- sample (transect) label column
- distance column

However, under most circumstances you will want to include other columns such as:

- stratum area
- sample effort (transect length) - for line transects
- angle - when radial distance and angle are measured
- cluster size - when objects are clusters

The columns should be separated by a delimiter (ASCII character), which can be either a tab, semicolon, comma or space. The order of the columns is not important, as you tell Distance which column is which during the import process. Each row should finish in a Carriage-return + Line-feed combination. This is the default end-of-line indicator used by most windows-based applications, so you usually don't have to worry about this.

While the order of the columns is not important, the order of the rows is. Before importing data into Distance, you should sort by stratum label (if you're importing more than one stratum), then sample (transect) label. This ensures that all data from the same strata are together, and within this all data from the same sample (transect) within strata are together. The order of observations within samples is not important.

The following is an example of a semicolon delimited input file. The first column is the stratum label, the next is the stratum area, then the transect label, transect length and finally the perpendicular distance.

```
Stratum A;100;Line 1A;10;14
Stratum A;100;Line 1A;10;8
Stratum A;100;Line 1A;10;22
Stratum A;100;Line 2A;10.3;7
Stratum A;100;Line 2A;10.3;37
Stratum A;100;Line 2A;10.3;13
Stratum B;123;Line 1B;5.7;
Stratum B;123;Line 2B;8.4;27
Stratum B;123;Line 2B;8.4;76
Stratum B;123;Line 2B;8.4;44
Stratum B;123;Line 2B;8.4;7
```

Notice that the record "Line 1B" has no distance in the final column - this is a transect where no objects were seen.

Notice also that all transects from the same stratum are grouped together, and all observations from the same transect are grouped together. If you accidentally forgot to sort the data before importing it, so that for example the first four lines looked like this:

```
Stratum A;100;Line 1A;10;14
Stratum A;100;Line 1A;10;8
Stratum A;100;Line 2A;10.3;7
Stratum A;100;Line 1A;10;22
```

then Distance would interpret this as three transects with labels "Line 1A", "Line 2A" and "Line 1A" again.

Distance treats each successive delimiter as indicating a new column. This means that you cannot import fixed-format data in the current version of Distance. Here's an example of some fixed-format data, where the stratum field is columns 1-10, stratum area is columns 10-20, etc.

```
Stratum_A      100    Line_1A   10     14
Stratum_A      100    Line_1A   10      8
```

If you wanted to import this data, you would have to find some way to delete the multiple spaces before importing it as space delimited:

```
Stratum_A 100 Line_1A 10 14
Stratum_A 100 Line_1A 10 8
```

In Distance, you are guided through the import process by the Import Data Wizard. This wizard can be started in one of two ways:

- from the last page of the Setup Project Wizard, by choosing the option **Proceed to Import Data Wizard**. This is the ideal way to import data into a new project.
- by selecting the menu item **Tools | Import Data Wizard**. This is the best way to add extra data from file into an existing project. You can also replace your existing data with the imported data - this is an option at the end of the Import Data Wizard.

Tip! You can streamline data import by having the columns in your text file in the same order as the fields in the Distance Data Explorer. That way, you can tick a box in the Import Data Wizard and have the wizard automatically pair up columns in the text file with fields in the Distance database. See also Streamlining import of data from one flat file.

Additional information about the Import Data Wizard is given in the Program Reference page Import Data Wizard. If you are having problems, check the page Troubleshooting the Import Data Wizard.

Tip! Once you have imported your data, you should always double-check that the correct number of strata, samples (transects) and observations are present in the Distance project. For example, if you forget to sort the data correctly (by stratum and sample label)

Warning! Importing large datasets into Distance takes a long time. We hope to improve the performance of the import routines in future releases.

Overview of More Advanced Import Topics

This section contains a brief mention of some more advanced uses of the Import Data Wizard. Further details of the wizard are in the Import Data Wizard section of the Program Reference. Some specific import scenarios are covered in the next topics after this one.

Importing A Subset of Layers

In the Import Data Wizard, you specify which layers you want to import to, and where the new data records should be located.

For example, this is useful for importing survey data where there is no stratum. By default, the stratum layer contains one record. So, in the Import Data Wizard, you specify that you want to import data into only the Sample and Observation layers, and that the new records should be added below the first record in the Stratum layer.

Another example where this is useful is when you already have one text file for each data layer, rather than one large text file containing all layers joined together.

Creating New Fields

The Import Data Wizard is capable of importing additional columns of data and will automatically create new fields in the Distance project file to hold them if the fields do not exist already. Examples are given later in the Users Guide where additional columns of data are useful - these data include extra columns in the Sample data layer (such as year of survey in multi-year surveys) and the Observation data layer (such as sex or species of animal).

Ignoring Columns

The wizard is also capable of ignoring columns in your data text file, so there is no need to exclude all unwanted columns from the file prior to import.

Appending New Data

You can choose whether to append the imported data to existing data, or to overwrite the existing data. This is useful, for example, for adding a new year of data to an existing project file.

Specifying Destination for New Records

You can specify where the new records will be added, relative to the parent layer, using the parent layer's ID or Label field.

For example, imagine you are adding a new year of survey data to an existing project file. You have covered new transects, so you want to add new records to the Sample layer, and you have new sightings for the Observation layer. You want the new transects to go into the correct place below their parent records in the Stratum layer.

To do this, you add a column to your text file containing the ID of the stratum for each new transect. You will also need a column saying which year the records come from – this should go in the Sample layer. In the Import Data Wizard you specify that you want to import into the Sample and Observation layer, and that you want to put new records below the ID field of the parent Stratum layer. You specify that the new data should be appended, and when you press the Finish button the new data are added to the old, in the correct stratum.

Importing Geographic Information

Unfortunately, there is currently no user-friendly way to import geographic information into Distance. The only way, apart from typing it in, is to create a shapefile in a separate package and then link this shapefile to the database – see Importing Existing GIS Data, below.

Streamlining import of data from one flat file

If you find yourself importing lots of data that are basically similar, then there are several steps you can take to make the import process quicker. This topic describes these steps, using an example of a single text file that contains all the survey data.

Before reading further, you should be somewhat familiar with the Import Data Wizard, and should read the Program Reference pages on the Import Data Wizard.

Let's consider the following data, stored in a text file:

```
Stratum A;100;Line 1A;10;14
Stratum A;100;Line 1A;10;8
Stratum A;100;Line 1A;10;22
Stratum A;100;Line 2A;10.3;7
Stratum A;100;Line 2A;10.3;37
Stratum A;100;Line 2A;10.3;13
Stratum B;123;Line 1B;5.7;
Stratum B;123;Line 2B;8.4;27
Stratum B;123;Line 2B;8.4;76
Stratum B;123;Line 2B;8.4;44
Stratum B;123;Line 2B;8.4;7
```

The data are semicolon delimited, and the columns are: stratum label, stratum area, transect label, transect length and distance. Normally, to get such data into a Distance project you would:

4. create a new project, going through the Setup Project Wizard, choosing the option to **Analyze a survey that has been completed**, and filling in the options in the successive screens.
5. Proceed to the Import Data Wizard, and specify the appropriate source file
6. In the Data File Structure screen of the Import Data Wizard, manually match up the field names with the columns in the text file.
7. Finish the Import Data Wizard and import the data.

This process can be made more efficient using some combination of the following tips:

- If you already have a project with the data structure you require, then in the first page of the Setup Project Wizard, choose the option to **Use an existing Distance project as a template.** See Creating a New Project in Chapter 4.
- Rather than manually match up field names with columns in the text file, make sure the columns are in the same order as the fields in the Distance database. Then, in the Data File Structure screen, tick the option **Columns are in the same order as they will appear in the data sheet**.
- Another alternative is to put the layer and field name in the first row of the text file. Then tick the option **First row contains layer names and field names of each column**. For example (the text below is supposed to be on one line, but may be wrapped on some screens/formats):

```
Region*Label;Region*Area;Transect*Label;Transect*Length;Observation
*Distance
Stratum A;100;Line 1A;10;14
Stratum A;100;Line 1A;10;8

. . .
```

The "*" delimiting layer names and field names in the code can be replaced by other delimiters. Alternatives are * | _ and . (.e., a full stop or period). During data import, you can choose from these alternatives the appropriate one for the text file you are importng.

Importing one file per data layer

Flat files, such as the one used in the example of the previous topic, are useful ways to store small datasets. However, for large datasets they are inefficient. For example, in the previous topic, the stratum label and area for stratum 1 was repeated six times. Imagine if there were 10,000 observations in stratum 1! A more efficient way to store and import large datasets is to have each data layer in a separate file, and to import one layer at a time.

Continuing the example from the previous topic, you would have 3 files:

File 1: stratum.txt
Columns: stratum label, area

```
Stratum A;100
Stratum B;200
```

File 2: transect.txt
Columns: stratum label, transect label, transect length

```
Stratum A;Line 1A;10
Stratum A;Line 2A;10.3
Stratum B;Line 1B;5.7
Stratum B;Line 2B;8.4
```

File 3: observation.txt
Columns: transect label, distance

```
Line 1A;8
Line 1A;22
Line 2A;7
Line 2A;37
Line 2A;13
Line 2B;27
Line 2B;76
```

```
Line 2B;44
Line 2B;7
```

Notice that the transect file contains a column giving the stratum of each transect, and that the observation file contains a column giving the transect of each observation. In general, each file has to have a column giving an unique identifier to the record in the parent layer. You don't need one for the stratum file because it's parent, the global layer, has only one record.

To import these data:

- Create a new distance project, with 4 data layers and appropriate fields – probably using the Setup Project Wizard.
- Begin by importing the stratum layer. Fire up the Import Data Wizard, and enter "`stratum.txt`" in the Data Source page.
- Under Data Destination, both the highest and lowest **Destination data layers** are the stratum layer (called Region by default). Under **Location of new records**, choose the first option **Add all new records under the first record in the parent data layer**.
- Under Data File Format, choose semicolon delimited, and in Data File Structure, match the columns in "stratum.txt" to the fields in the stratum layer. Click **Next** and **Finish**.
- Use the Data Explorer to check the stratum data were imported correctly.
- Now import the transect file. This time in Data Destination, the highest and lowest **Destination data layers** are the transect layer. Under **Location of new records**, choose the second option, **Input file contains a column corresponding to the following field in the parent data layer**, and make sure the label field is selected from the drop down box.
- In the Data File Structure page, match the columns in "transect.txt" to those in the Distance database, including the stratum label field.
- Import the data, and check it in the Data Explorer.
- Repeat this process for the observation data file.

Non-unique label fields

If the label field is not unique, then you will have to add an extra column containing the ID of each record. For example, imagine that the transect labels are not Line 1A, Line 2A, Line 1B, Line 2B, but instead are Line 1, Line 2, Line 1, Line 2. In this case the observation data file will need to be as follows:

File 3: observation.txt
Columns: transect ID, distance

```
1;8
1;22
2;7
2;37
2;13
4;27
4;76
4;44
4;7
```

When the transects are created, they are assigned IDs sequentially, so transect "Line 1" in stratum A will have ID 1, transect "Line 2" in stratum A will have ID 2, "Line 1" in stratum B will be ID 3, and "Line 2" in stratum B will be ID 3. In the above file, because the transect labels are not unique, the transect IDs have been used instead. The only difference in the Import Data Wizard will come in the Data Destination step, where under **Location of new records**, the **Field name** will be "ID", rather than "Label".

Geographic (GIS) Data

Distance has a built-in GIS (Geographic Information System), which allows it to store and manipulate spatially-referenced data. Geographic data for spatially-reference data layers is stored in a shapefile, a widely used data format invented by the GIS company ESRI (see Chapter 4 - Distance Projects). Geographic data is used in the survey design part of Distance, and in future releases will be used in analyses involving spatial modelling of density.

There are two types of Distance project: geographic projects, which can contain spatial data, and non-geographic projects, which can not. The project type can be set when the project is created (see Setup Project Wizard in the Program Reference). To see whether a project is geographic, look under the Geographic tab of the Project Properties dialog (choose **File** | **Project Properties…**).

Not all data layers in a geographic project have to be spatially-referenced, although all are by default. This is set when the layer is created. To see whether a data layer is spatially-referenced, in the Data Explorer, either (i) look in the data sheet for a column with name "Shape", or (ii) look under the Geographic tab of the Layer Properties dialog (choose **Data** | **Data Layer Properties…**).

Shapes in a spatially-referenced data layer can be one of three types: points, lines or polygons. The shape type is set when the layer is created. You can see the shape type of a layer, either by looking under the "Shape" field in the data sheet, where the shape type is written for each record, or by looking under the Geographic tab of the Layer Properties dialog.

Geographic data in a data layer can be stored according to a geo-coordinate system, and projected for viewing and survey design calculations. This is covered in the section on Coordinate Systems and Projections.

There are two ways to get geographic data into distance: type it in using the Shape Properties Dialog (see Shape Properties Dialog), or import it (see Importing Existing GIS Data).

If you are experiencing strange behaviour or error messages after importing GIS data into Distance, it could be because of geometry problems in the data. For more on this see GIS Problems in Chapter 12.

Viewing and Manipulating Geographic Data

Viewing Geographic Data in Maps

Geographic data in Distance can be viewed in maps, which are accessable from the **Map Browser** – accessed via the **Maps** tab of the **Project Browser**. The Map Browser allows creation of new maps, and allows you to sort, rename, delete and preview the maps that have been created. For more information see Map Browser in the Program Reference.

From the Map Browser, select a map and choose **View Map** to open the map in a Map window. The map window allows you to add or remove data layers, pan and zoom, and view information about features on the map (using "Map Tips"). In the future it will be possible to customize the way each data layer is presented – the colour, etc of each shape – but this facility is currently not developed. For more information, see Map Window in the Program Reference.

Maps are also produced during survey design, when using a design to investigate probability of coverage and to produce example surveys. These maps are displayed in the Results pages of the Survey Details and Design Details windows. For more information, see Chapter 6 - Survey Design in Distance.

Tip! Maps in the results pages of Survey Details and Design Details cannot be customized, but copies can be saved to the Map Browser for customization by pressing the **Add to Map Browser** button .

Viewing and Manipulating Geographic Data in the Shape Properties Dialog

If a data layer contains geographic information, it will have a field called "Shape" in the Data Explorer. The contents of this field will depend on the type of shape – either "Point", "Line" or "Polygon". Double-clicking on any of these will open the Shape Properties Dialog, which lists the vertices of the shape and allows you to edit them. For more information, see the Shape Properties Dialog page of the Program Reference. You can also use the Shape Properties Dialog to import shapes from a text file or spreadsheet – for more information, see Importing GIS Data via the Windows Clipboard, later in this chapter.

Coordinate Systems and Projections

Advanced Topic

This section provides a brief introduction to the use of coordinate systems in Distance. For more information, refer to any good book on cartography.

About coordinate systems

There are two types of coordinate systems: geographic and projected. Geographic coordinate systems use latitude and longitude coordinates to define the position of a point, line or polygon on the earth's three-dimensional surface. Most geographic data is stored in latitude and longitude, according to a specified geo-coordinate system. Projected coordinate systems use a mathematical conversion to transform latitude and longitude coordinates to a two-dimensional surface. Most maps use a projected coordinate system, and calculations such as distance and area are performed on projected data.

Geographic coordinate systems

Unfortunately for us, the earth is not a perfect sphere, or even a perfect ellipsoid. Instead, it is nearly ellipsoidal and in addition is covered in lots of small lumps and bumps. Geographic coordinate systems approximate the shape of the earth using a reference sphere or ellipsoid. The accuracy of the approximation depends upon which coordinate system you use and which part of the earth you use it for.

Projected coordinate systems

Because the earth is round and maps are flat, getting information from a curved surface to a flat one involves a mathematical formula called a map projection or simply a projection

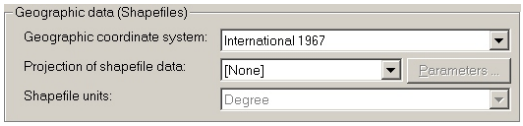
This process of flattening the earth will cause distortions in one or more of the following spatial properties: distance , area, shape, direction.

No projection can preserve all these properties; as a result, all flat maps are distorted to some degree. Over a small geographic area the distortions caused by the map projection are not significant, but the selection of an appropriate map projection is crucial for larger regions. Fortunately, you can choose from many different map projections. Each is distinguished by its suitability for representing a particular portion and amount of the earth's surface and by its ability to preserve distance, area, shape, or direction. Some map projections minimize distortion in one property at the expense of another, while others strive to balance the overall distortion.

Coordinate systems and Distance data

Distance can cope with data stored in either a geographic or projected coordinate system, or neither (sometimes referred to as "non-earth" – in the Distance interface we use "[None]"). In Distance, a projected coordinate system is defined as a geographic coordinate system plus a projection together with any projection parameters required.

The default coordinate system for geographic data in a project is set in the Geographic tab of the Project Properties dialog.



Data section of the Geographic tab, Project Properties dialog

You set the coordinate system of a data layer when it is created – by default it is the same as the default coordinate system, but this is not required. (The only requirement is that all data layers use the same datum.)

Do you need to worry about the coordinate system of the data?

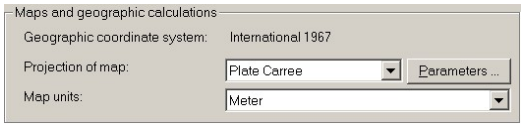
Most geographic data is stored as latitude and longitude according to some geographic coordinate system. Latitude and longitude are expressed in angular units (usually decimal degrees). If you want to work with survey design, you will likely want to work in linear units (e.g., meters), so you will need to transform your data. To do this you will need to know the geocoordinate system.

If your data are already expressed in linear units, then likely they are already projected. This can happen, for example, because you digitized the study area using a map. So long as you are happy with the projection, then you can set the geographic coordinate system to "[None]" and forget about coordinate systems.

Similarly, if your study area is small, and you have measured its boundaries directly, then no coordinate system is required.

Coordinate systems, maps and calculations in Distance

If the data are stored in a geographic coordinate system then you can project the data for viewing it on maps and performing calculations. All maps must use the same projection, which is set in the Geographic tab of the Project properties dialog.



Maps section of the Geographic tab, Project Properties dialog

If the data are stored projected, then this projection is used for all maps and calculations, while if the data have no coordinate system then they cannot be projected.

In survey design, a different projection can be defined for each design, so you can compare the effect of different projections on the results. (See Chapter 6 - Survey Design in Distance for more on survey design.)

Tip! Projecting data takes time. This can significantly affect the performance of survey design calculations, such as calculating coverage probability and creating new surveys. Therefore, if you always use the same projection for a particular set of data, consider projecting the data using an

external GIS package and storing the projected data in a shapefile. When you bring the data into Distance, set it as no coordinate system (you can do this by setting the coordinate system and projection to **[None]** in **File** | **Project Properties** | **Geographic**, and setting the units correctly, and then not copying over the .prj part of the shapefiles when you import your shapefiles – see Importing Existing GIS data). Distance will then treat the data as simple x,y coordinates and will not spend time projecting and re-projecting it.

Which projection?

Over a small study area the projection used will make relatively little difference. Over larger areas the projection can make a significant difference.

If you need to project your data but are not sure which projection to use, probably the best option is to cheat and refer to maps of your study area to see what projection they use (the map will usually give the projected coordinate system – a literature search will reveal the composite projection, projection parameters and geocoordinate system).

Alternatively, many cartography books describe the properties of the different projections, which may help you decide which is appropriate. You will need to consider the following questions:

- Which spatial properties do you want to preserve? Is it just for displaying the study area, or for performing calculations such as for survey design, transect distance, etc.
- Where is the study area? Is your data in a polar region? An equatorial region?
- What shape is the study area? Is it square? Is it wider in the east–west direction?
- How big is the study area?

Map projection classifications

Map projections can be generally classified according to what spatial attribute they preserve.

- *Equal Area* projections preserve area. Many thematic maps use an equal area projection. Maps of the United States commonly use the Albers Equal Area Conic projection.
- *Conformal* projections preserve shape and are useful for navigational charts and weather maps. Shape is preserved for small areas, but the shape of a large area such as a continent will be significantly distorted. The Lambert Conformal Conic and Mercator projections are common conformal projections.
- *Equidistant* projections preserve distances, but no projection can preserve distances from all points to all other points. Instead, distance can be held true from one point (or a few points) to all other points or along all meridians or parallels. If you will be using your map to find features that are within a certain distance of other features, you should use an equidistant map projection.
- *Azimuthal* projections preserve direction from one point to all other points. This quality can be combined with equal area, conformal, and equidistant projections, as in the Lambert Equal Area Azimuthal and the Azimuthal Equidistant projections.

Other projections minimize overall distortion but don't preserve any of the four spatial properties of area, shape, distance, and direction. The Robinson projection, for example, is neither equal area nor conformal but is aesthetically pleasing and useful for general mapping.

GIS Data Format

Geographic data is stored in Distance as ESRI shapefiles – a standard, vector-based GIS format. There are several ways to get GIS data into Distance (see Importing Existing GIS Data). If you have a GIS, you can probably simply export data from there into a shapefile format, and use the instructions in the above section to get the data into Distance. If not, you will probably need to enter the geographic coordinates into a text file, database file or the Distance Shape Properties Dialog by hand.

For point and line data, it is quite straightforward to see which order to enter the points, or the vertices of the lines. For polygons, especially complex polygons (for example those containing holes), it is less straightforward. If you are entering data by hand, it is worth first checking the ESRI Shapefile Technical Description – available from the ESRI web site and also from the Program Distance Web Site under "Support, Updates and Extras". An excerpt of this information is given below, under GIS Format for Polygons.

For more information about checking a shapefile is valid, see GIS problems in Chapter 12.

GIS Format for Polygons

The following is an excerpt from the ESRI Shapefile Technical Description. The full text is available from the Program Distance Web Site under "Support, Updates and Extras".

A polygon consists of one or more rings. A ring is a connected sequence of four or more points that form a closed, non-self-intersecting loop. A polygon may contain multiple outer rings. The order of vertices or orientation for a ring indicates which side of the ring is the interior of the polygon. The neighborhood to the right of an observer walking along the ring in vertex order is the neighborhood inside the polygon. Vertices of rings defining holes in polygons are in a counterclockwise direction. Vertices for a single, ringed polygon are, therefore, always in clockwise order. The rings of a polygon are referred to as its parts.

Because this specification does not forbid consecutive points with identical coordinates, shapefile readers must handle such cases. On the other hand, the degenerate, zero length or zero area parts that might result are not allowed.

[…]

The following are important notes about Polygon shapes.

- The rings are closed (the first and last vertex of a ring MUST be the same).
- The order of rings in the points array is not significant.
- Polygons stored in a shapefile must be clean. A clean polygon is one that

    1. Has no self-intersections. This means that a segment belonging to one ring may not intersect a segment belonging to another ring. The rings of a polygon can touch each other at vertices but not along segments. Colinear segments are considered intersecting.
    2. Has the inside of the polygon on the "correct" side of the line that defines it. The neighborhood to the right of an observer walking along the ring in vertex order is the inside of the polygon. Vertices for a single, ringed polygon are, therefore, always in clockwise order. Rings defining holes in these polygons have a counterclockwise orientation. "Dirty"

polygons occur when the rings that define holes in the polygon also go clockwise, which causes overlapping interiors.

Importing Existing GIS Data

There are four ways to get geographic information into Distance:

- Enter the vertices (corners) of each shape by hand using the Shape Properties Dialog (see Shape Properties Dialog in the Program Reference). Clearly, this is only useful for a small number of simple shapes.
- Copy the vertices of each shape from a text file or spreadsheet and paste them into the Shape Properties Dialog. This option will work well if you have only a few shapes to import (such as just a few regions) and already have the vertices in some other file.
- Copy an existing ESRI shapefile into the Distance project's Data Folder. This will work well if you already have the geographic data in a shapefile (or can export your data into this format), and have access to a GIS package for preparing the shapefile.
- Link and existing ESRI shapefile by editing the Data File. This requires a separate package for preparing the shapefile, and Microsoft Access for editing the Data File.

The last three options are covered in more detail in the sections below. Before reading further, you need to understand how data, and particularly geographic data, are stored in Distance. Make sure you have read all of Chapter 5 - Data in Distance up to this point, and also read the section How Distance Stores Data, in the Inside Distance appendix.

Consider projecting your data before importing it, and then importing it into Distance without the projection – see the tip in Coordinate Systems, Maps and Calculations in Distance on page 14 for details.

Importing GIS Data via the Windows Clipboard

The Shape Properties Dialog has a facility to copy and paste the vertices (corners) of an individual shape to and from the Windows clipboard. You can use this to transfer GIS data between Distance and other formats such as text files and spreadsheets. For a step-by-step example of importing GIS data from a text file into Distance, see the Getting Started chapter Example 3: Using Distance to Design a Survey.

To copy data from a spreadsheet or text file into Distance:

- Highlight the data in the text file or spreadsheet and copy it to the Windows clipboard.
- In the Distance project you want to copy to, select the shape you wish to replace in the Data Explorer and double-click on it to open the Shape Properties Dialog.
- Choose **Paste from Clipboard**.

To copy data from Distance to a spreadsheet or text file:

- In Distance, double-click on the shape in the Data Explorer. This opens the Shape Properties Dialog.
- Choose **Copy to Clipboard**

- In your text file or spreadsheet editor, select the area you want to paste to and choose Paste.

Text file format

Each vertex should be on a separate line, with the x coordinate first, followed by a tab, followed by the y-coordinate. For example, a text file containing the following four lines indicates a square:

```
0       0
0       100
100     100
100     0
```

To separate the parts of a multi-part polygon, leave a blank line (or have a line that contains anything other than the above number-tab-number format). For example the following indicates two triangles:

```
0       0
0       100
100     0

100     0
100     100
200     0
```

Spreadsheet format

Each vertex should be in a separate row, with two columns: the first for the x-coordinate and the second for the y-coordinate. To separate the parts of a multi-part polygon, leave a row blank.

	A	B
1	0	0
2	0	100
3	100	0
4		
5	100	0
6	100	100
7	200	0

Example from an Excel spreadsheet, showing data for two triangles. Both columns are highlighted, ready to copy to the windows clipboard.

Importing GIS Data by Copying an Existing Shapefile into the Data Folder

In Distance, each geographic data layer has an associated shapefile. By default, these shapefiles are located in the Data Folder for that project. For example, in the Mexico sample project, the data layer "Mex" has an associated shapefile "Mex.shp" (and other files "Mex.dbf", "Mex.shx" and "Mex.prj"). In this method, we import the GIS data by copying our shapefiles into the Data Folder, and then renaming them so they overwrite shapefiles already attached to a data layer in Distance. That way, next time Distance opens the project, it will use our shapefiles rather than the original ones. The advantage of this method is that it does not require the use of a database package to edit the project's Data File (see Importing GIS Data by Linking and Existing Shapefile to the Project Data Folder, below). The disadvantage is that your shapefile has to be copied into the Data Folder from its original location – so you then have two copies of your GIS data to manage.

The following instructions assume that you have software to create and edit shapefiles (e.g., ESRI ArcView). They guide you through the process of importing a single shapefile to and linking it to a single Distance data layer. If you are importing more than one shapefile then you need to repeat the same process for each shapefile. They start by assuming you have created the Distance project, and have the shapefile you wish to import. The Distance project must be geographic.

Prepare the Distance Data Layer

- In Distance, open the project file.
- If you have not yet created the data layer that will hold the shapefile, then you should do so now.
    - In the Data Explorer, click on the **Create New Data Layer** button. The **Create New Layer** dialog box opens.
    - Give the layer an appropriate name, and choose the appropriate parent layer and layer type. Make sure that the options **Create new tables for layer**, **Create internal data table** within project file and **Create shapefile** are all ticked.
    - Note down the name of the shapefile that will be created. For example, if the new data layer is called "Antarctica", the shapefile will be "Antarcti.shp"
    - Click the **OK** button. The new layer will be created.
- If you had already created the data layer that will hold the shapefile, open the **Data Layer Properties** dialog for that layer (highlight the layer in the **Data Explorer**, and click the **Data Layer Properties** button). In the **Geographic data** tab, note down the name of the shapefile (first line).
- Make sure that the data layer contains the same number of records as the shapefile. For example, if your shapefile contains 5 shapes (corresponding, say, to 5 strata), then the data layer inside Distance needs to have 5 records, with ID 1 to 5.

  If the data layer does not contain the same number of records (for example, because you have just created it and it is empty), then add the appropriate number of new records. (See the Data Explorer section of the Program Reference for how to do this, if you do not know.)
- Close the project in Distance.

Prepare the Shapefile

- In Windows, go to the Data Folder for this project. Locate the shapefile currently associated with the data layer you are going to import to – in the above example this was called "Antarcti.shp". Delete this file, and all other related files – those with the same name but ending in ".dbf", ".shx", ".prj", etc.
- Locate the shapefile that you have previously created, and wish to import. Copy this shapefile (and all associated files) into the Data Folder. Rename the file (and all associated files) so they have the same name as the shapefile you just deleted. In the above example, we would rename our shapefile so it is called "Antarcti.shp", "Antarcti.dbf", etc.
- Open the newly renamed shapefile in your GIS package, and add a field to the table. The field should be able to contain long integer numbers – in ArcView this means the field type should be "Number", the width 16 and decimal places 0. Name this field "LinkID".
- The LinkID field will be used to link records in the shapefile to records in Distance's internal data table for this layer. Each LinkID value must correspond with a value in the ID field of Distance's internal table.
    - If you only just created the Data Layer in Distance, then it doesn't matter what order you number the records in the

LinkID field of the shapefile. Start with 1, and go up to the number of records you have.

- If you created the Data Layer previously in Distance, and already have information in the layer such as Labels, Areas, etc., then you should take care to number the records in the shapefile so that the LinkID field of a shape corresponds to the ID value of that record inside Distance.

- Once you have added records to the LinkID field, you can close the shapefile and exit your GIS.

Clean up

- You can now reopen the project inside Distance. The new shapefile should now be attached. You can confirm this by looking in the **Data Layer Properties**, or by creating a new **Map**.

- If your data are from a specific coordinate system, and you did not copy across .prj (projection) file with the rest of the shapefile, then you need to tell Distance about the coordinate system. You do this in the Geographic tab of the Data Layer Properties.

Importing GIS Data by Linking and Existing Shapefile to the Project Data Folder

Advanced Topic

This method is similar to the previous one, but does not require you to copy the shapefile into the Data Folder, or to rename it. The advantage, therefore, is that you only have one copy of your shapefile to manage. The disadvantages are:

- it is more complicated, requiring you to edit the Data File using a database package

- because the shapefile remains outside the Data Folder, it is harder to move the project onto other machines (for example, **Export Project** only copies files in the Data Folder).

The following instructions assume that you have software to (1) create and edit shapefiles (e.g., ESRI ArcView), (2) edit the Distance Data File DistData.mdb (e.g., Microsoft Access 97). If you have Access 2000 or later, then you should read Accessing DistData.mdb using newer versions of Access in the Appendix – Inside Distance.

Prepare the Distance Data Layer

- Follow the instructions under this section in Method 1:Prepare the Distance Data Layer.

Prepare the Shapefile

- Open the shapefile you wish to import in your GIS package, and add a field to the table. The field should be able to contain long integer numbers – in ArcView this means the field type should be "Number", the width 16 and decimal places 0. Name this field "LinkID".

- The LinkID field will be used to link records in the shapefile to records in Distance's internal data table for this layer. Each LinkID value must correspond with a value in the ID field of Distance's internal table.

  - If you haven't created the Data Layer in Distance yet, then it doesn't matter what order you number the records in the

LinkID field of the shapefile. Start with 1, and go up to the number of records you have.

- If you have created the Data Layer already in Distance, and already have information in the layer such as Labels, Areas, etc., then you should take care to number the records in the shapefile so that the LinkID field of a shape corresponds to the ID value of that record inside Distance.

- Once you have added records to the LinkID field, you can close the shapefile and exit your GIS.

Edit the Distance Data File

- In your database package, open the project's Data File, DistData.mdb. This is located in the project's data folder.
- Open the DataTables table. Locate the record that corresponds to the shapefile of the data layer in which you are interested. For example, if you have created a data layer called "Antarctica" then look for the record with a LayerName of "Antarctica" and TableName that starts with "geo" (e.g., "geoAntarcti"). The SourceDatabaseType should be "Geog".
  - For this record, change the SourceTableName to the name of the shapefile you want to import. Don't use any suffix. For example, if the shapefile files are "Boundary.shp", "Boundary.shx" and "Boundary.dbf" then you enter "Boundary".
  - Change the SourceDatabaseName to the path of the folder containing your shapefile. You must include the full absolute path – e.g., "D:\data\shapefiles". If the shapefile is located in the Data Folder for this project (for example if you copied it there), then this field should be blank.
- Close the database DistData.mdb

Clean up

- You can now reopen the project inside Distance. The new shapefile should now be attached. You can confirm this by looking in the **Data Layer Properties**, or by creating a new **Map**.
- If your data are from a specific coordinate system, and you did not copy across .prj (projection) file with the rest of the shapefile, then you need to tell Distance about the coordinate system. You do this in the **Geographic** tab of the **Data Layer Properties**.
- If the data imported correctly, you can delete the old shapefile from the Data Folder.

Advanced Data Topics

Linking to Data From Other Databases

It is possible to directly link to data in external database tables, spreadsheets and text files, instead of importing them into the distance database. This is an advanced technique, and should be used only by those confident poking around inside Microsoft Access databases. The technique involves editing the project Data File (which is a Microsoft Access database) using Access, or some other

database tool (i.e., outside of the Distance interface). Note that the Distance database engine is quite unforgiving - if you make a small mistake specifying the location of the data, Distance will generate an error next time you try to open the project within Distance.

For more information, see Linking to External Data from Distdata.mdb in the Inside Distance Appendix. The text there refers to a sample project, LinkingExample.dst, which is located in the sample projects folder (see Sample Projects).

Chapter 6 - Survey Design in Distance

Introduction to Survey Design in Distance

This chapter gives some background information about survey design, and an overview of the interface for survey design in Distance. The methods implemented here are based on work by Strindberg (2001), and are described in Strindberg et al. (2004). This latter text, which is Chapter 7 of Buckland et al. (2004) is recommended reading for anyone using the design engine in Distance. In addition, Thomas et al. (2007) review these concepts, with particular application to design of shipboard surveys in complex regions using Distance.

A number of frequently used survey designs are implemented within the geographic survey design component of the Distance software (see Design Classes Available in Distance, below). You can use the component either to evaluate the properties of a design class or to generate an instance from that class which can act as a survey plan. Simulation is used to calculate design class properties, such as coverage probability, estimates of distance travelled while on- and off- effort and between sampling locations. Design classes comprise a sampler type, e.g. point or line, and a design type associated with the sampler. The design type has an associated survey design algorithm, which has been automated to generate the survey designs.

By creating a number of different designs in Distance, you can compare the properties of the designs using simulation (see previous paragraph), and then select a suitable design for your study. Most designs are stratified, in which case the different designs may contain different number of strata or stratum boundaries.

Within each stratum, sampler points, or lines, might be randomly located. Alternatively, they may be placed on a regular grid that is randomly superimposed on the stratum. Sometimes, more complex algorithms are required. For example, shipboard surveys typically use continuous zigzag samplers, so that costly ship time is not wasted in travelling from one line to the next. A number of different zigzag designs are implemented in the software. When these designs are realized within a convex survey region the sampler line is continuous. Non-convex survey regions lead to some sampler discontinuity, as the sampler is clipped against the survey region boundary.

For complex surveys, in which coverage probability is not uniform, the software permits evaluation of coverage probability by location, using simulation as mentioned above. Transect survey data are frequently analysed under the assumption of an equal coverage probability design, as this avoids the necessity of making assumptions about the distribution of the survey population. A design

that leads to uneven coverage probability throughout the survey region can lead to biased abundance estimates, if the analysis assumes coverage probability is constant. Unbiased estimates can be calculated from the sample data if an appropriate estimator, such as the Horvitz-Thompson-like estimator, is used and the unequal coverage probabilities are taken into account. The properties that such an algorithm should possess include the following: no part of the survey region should have zero coverage probability, and coverage probability should be as nearly constant as possible. The former property affects bias, whereas the latter affects efficiency. We hope to implement an unequal coverage estimator in Distance in the future.

One advantage of design automation by means of software is that it enables each possible design to be compared for efficiency, accuracy, and bias of the subsequent abundance estimates, using simulation over a population of interest. Simulation capabilities of this nature are planned for a future version of Distance.

An introduction to survey design issues for distance sampling is given in Chapter 7 of Buckland et al. (1993, 2001). These issues will be covered in detail in Strindberg (in prep.) and Buckland et al. (in prep.). A good general text on survey design is Thompson (1992).

Design Classes Available in Distance

The following design classes are available in Distance. We expect to add more in future releases. Generally, all except the zigzag designs produce even probability of coverage in the survey region (or stratum, for stratified surveys), although care should be taken at the region boundaries (see Concept: Edge Effects). For more on zig-zag designs, see Concept: Zigzag Sampling Designs.

Sampler	Design Class	Example	Description
Point	Simple Random Sampling		Randomly distributes a fixed number of points over the survey region.
	Systematic Grid Sampling		Randomly superimposes a systematic point grid of fixed dimensions and rotation onto the survey region.
Line	Parallel Random Sampling		Randomly distributes a number of parallel lines over the survey region.
	Systematic Random Sampling		Randomly superimposes a systematic set of parallel lines onto the survey region.
	Systematic Segmented Trackline Sampling		Randomly superimposes a systematic set of segmented parallel lines onto the survey region. A set of parallel tracklines is used for this purpose.
	Systematic Segmented Grid Sampling		Randomly superimposes a systematic set of segmented parallel lines onto the survey region. A set of grid points is used for this purpose.
	Equal Angle Zigzag		Superimposes a continuous zigzag sampler of fixed angle on the survey region.
	Equal Spaced Zigzag		Superimposes a continuous zigzag sampler that passes through equally spaced points on opposite sides of the survey region boundary.
	Adjusted Angle Zigzag		Superimposes a continuous zigzag sampler whose angle is continuously

			adjusted by survey region height.

Survey Design Concepts

Concept: Coverage Probability

The coverage (or inclusion) probability at an arbitrary location within the survey region is the probability of it falling within the sampled portion of the survey region. Transect survey data are frequently analysed under the assumption of an equal coverage probability design. A design with uneven coverage probability leads to biased abundance estimates if even coverage probability is assumed.

Thus, in some respects the ideal is to attain equal probability of coverage throughout the survey region, as this simplifies the statistical analysis. However, if equal coverage probability is not feasible then it is possible to use a sampling design that gives different, but known coverage probabilities throughout the survey region; unbiased estimates can be calculated from the sample data if an appropriate estimator, such as a Horvitz-Thompson estimator, is used and the unequal coverage probabilities are taken into account. (A generalized Horvitz-Thompson-like estimator is planned for a future version of Distance.) Even if an estimator that takes unequal coverage probabilities into account is used, designs providing nearly even coverage probabilities are preferable. Animals detected in a region of relatively low coverage probability can contribute substantially to the abundance estimate, and estimation may be very imprecise.

It is also more difficult to get precise estimates of the coverage probabilities if these probabilities are small. A high coverage probability leads to more robust estimation, and improved precision for the coverage probability estimates, but this requires an increase in effort, which may not be affordable.

Simulation can be used to estimate the coverage probability at locations throughout the survey region for those designs that give uneven coverage probability. Even for the more straightforward designs that are frequently assumed to give even coverage probability throughout the survey region, simulations can be used to examine the potential problems caused by edge effects at the survey region boundaries.

For more details about the various options associated with estimating coverage probability, see the Program Reference section on Coverage Probability Information on the Design Results Tab.

Concept: Edge Effects

Point or line samplers have an associated radius or width, so parts of the areas sampled by each point or line sampler may fall outside the survey region. If the relative area of the sampled region is small relative to that of the survey region then discarding sampling units that intersect the boundary of the survey region causes negligible bias. However, if the relative area is large then discarding sampling units at the edge of the survey region may cause considerable bias in the estimates.

Minus Sampling

If points or lines are generated exclusively within the survey region we call this minus sampling. This leads to some under-sampling at the edge and an uneven coverage probability. Depending on the relative area of the sampled region and the relative density of the population along the edge of the survey region, this may or may not lead to significant bias in the estimates. This under-sampling is due to potential line or point samplers that lie just outside the boundary, whose sampled area intersects the survey region.

Plus Sampling

If a buffer zone is created around the survey region, then samplers can be generated in the survey region plus the buffer zone, which we call plus sampling. Plus sampling leads to an even coverage probability, but to some loss in efficiency (as part of the survey effort falls outside the region of interest).

With either plus or minus sampling, portions of the sampled area associated with samplers along the edge of the survey region will fall outside the region itself. If the terrain immediately outside the survey region differs substantially from that within the region and an abundance estimate for a particular habitat is required, then you should not include sample data from those areas outside the survey region. On the other hand, if the buffer zone surrounding the survey region is such that no population members are found in it (e.g. the survey region is an island the buffer the surrounding water), then the abundance estimate can be obtained for the survey region together with the buffer region and the same abundance estimate will apply to the survey region.

Concept: Zigzag Sampling Designs

Zigzag line transect designs can be more efficient than conventional parallel line designs, because no time is spent moving from one line to the next ("off effort"). This type of design is often used in shipboard surveys, where ship time is extremely costly and survey areas are large, so moving from one line to the next in a conventional design would be very expensive.

The downside is that zigzag designs zigzag surveys are difficult to generate in complex survey regions (see Zigzag Sampling in Non-convex Regions). Further, some zigzag designs do not produce even probability of coverage for anything but a rectangular survey design.

Here, we give a brief overview of the zigzag designs available in Distance – for more information see Strindberg (2001), Stringberg et al. (2004) and Strindberg and Buckland (2004).

The lines in a zigzag design are generated with respect to a **design axis**, which is a user-defined line overlaid on the survey region.

Distance offers 3 different zigzag design classes:

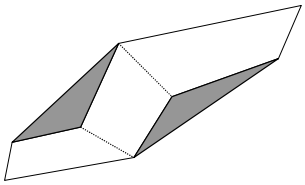
- **Equal Angle Zigzag**. In this design class, the angle of the lines is fixed with respect to the design axis. This design class produces even coverage only if the survey region is rectangular, and then only if the design axis runs parallel to one side of the rectangle.
- **Equal Spaced Zigzag**. Here, the lines run through equally spaced points on opposite sides of the survey region boundary. This design class also produces uneven coverage probability for all but rectangular survey designs. However, it gives more even coverage probabilities than the equal angle design. The coverage probabilities become more even as you increase the sampling intensity (i.e. line length), so you have to make a trade-off between the length of line you can afford to survey and the evenness of the coverage probability. There is also an issue with how to place the first and last lines – see First and last line placement in equal spacing zigzag designs.
- **Adjusted Angle Zigzag**. In this design class, the angle of the lines is continuously adjusted, depending on the height of the survey region. If the survey region is convex, this produces a design with even probability of coverage in the direction of the design axis. The downside is that it is difficult to implement on the

ground – in practice you would approximate the design using small straight segments.

Zigzag Sampling in Non-convex Regions

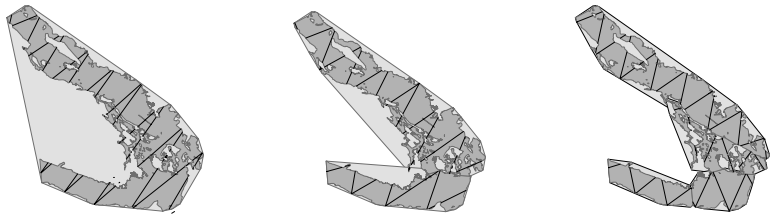
Zigzag sampling designs can only be generated in a convex survey region. If any of the survey strata in the survey layer are non-convex, then you can choose to generate the design for each stratum in either a **convex hull** or **bounding rectangle**. For non-convex strata, the zigzag line will no longer be continuous. The amount of discontinuity is generally less using the convex hull, but this may lead to uneven coverage probabilities. If simulation shows such an effect to be extreme, then it's better to use the bounding rectangle. Details of how to set these options are given in the Program Reference topic, Zigzag Sampling Non-convex Survey Region Options.

Another way of dealing with non-convex survey regions is by using stratification. The figure below shows how you can make a non-convex region convex by defining 3 strata. This is only an option for certain survey regions whose size and shape permit such stratification.



Stratifying the survey region into 3 strata can eliminate non-convexity or at least reduce the discontinuity in the sampler

Even if stratification doesn't let you take care of non-convexity entirely, it may at least reduce the discontinuity in your sampler. An example of this, for a real survey in British Colombia, is shown below.

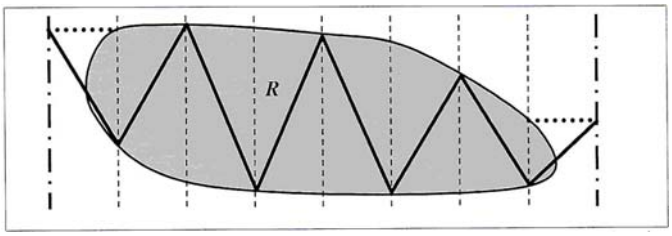


Single realizations of three equal-spacing zigzag designs applied to inshore waters of southern British Colombia (taken from Fig 3. of Thomas et al. 2007; see that paper for more details). In the first, the study region is treated as a single stratum. In the second it is divided into two strata. In the third it is divided into four strata. The designs are laid out in covex hulls fit to the strata; note that as the number of strata goes up, the strata become more convex and so the amount of discontinuity (i.e., off-effort time) between lines decreases.

First and last line placement in equal spacing zigzag designs

The material in this section is a brief abstract of Strindberg (2001), section 4.6, which should be consulted for full details on the algorithm used.

The equal spacing zigzag design works by placing a series of equally spaced lines across the study area, perpendicular to the design axis, and then joining up alternate waypoints formed by the intersection of these lines with the survey area boundary. However, for the first and last transect lines, the equally spaced perpendicular lines do not intersect the survey area boundary. Instead, a line is drawn from the survey area boundary until it intersects the perpendicular line at ninety degrees (the dashed line in the figure below), and this is used as the waypoint.



Method of constructing the first and last lines in an equal spacing zigzag design. Reproduced from Figure 4.6 of Strindberg 2001.

Although this method does not produce exactly equal coverage in the area of the first and last transects, it usually comes close (see, e.g., Thomas et al. 2007). Alternative algorithms may be implemented in future, such as the use of an adjusted angle zigzag sampler for the first and last segments.

Setting Up a New Project for Survey Design

This section outlines the options for creating and setting up a new project ready for use with the survey design component of Distance. You should read Chapter 4 - Distance Projects and Chapter 5 - Data in Distance before going any further!

Creating the new project

The first step is to create a new Distance project, using **File | New Project…**. In the first page of the New Project Setup Wizard, choose the option **Design a new survey**. Click **Next** and on the next page click **Finish**.

Importing or entering geographic data

If you've followed the instructions above, Distance has created a new project which contains a geographic global data layer with one record. In many cases, your survey designs will include strata. If this is true for you, then create a new layer below the global layer, and make it of type Stratum (see Chapter 5 Changing the Data Structure). Then add one record to the new layer for each stratum.

You're now ready to import or enter your data – see Chapter 5 Getting Data into Distance, for instructions how to do this. The most common route is to import the data by copying the appropriate shapefiles into the project folder. If you take this route, note that you need shapefiles both for your stratum layer(s) and the global layer. It is easy to form the global polygon in your GIS by joining the polygons from the stratum shapefile. You need to have a valid shapefile for the global layer before you can create a coverage probability grid (see below), even if you are doing the design at the stratum layer.

Creating a coverage layer

The next step is to create a grid of points at which coverage probability will be assessed for your designs – this is called a coverage layer.

In the Data Explorer, click on the icon for the global layer in the Data Layers tree (), and then click the **Create New Data Layer…** button. Under "Layer type", choose **Coverage**, and then click **Properties…** and fill in the required grid spacing and other properties. Once you have the options set the way you want, click **OK** in the Grid Properties dialog, and then **OK** again in the Create New Layer dialog. Distance will then create the coverage layer and the points that go in it.

You can check the number of points from the Data Explorer by clicking on the icon for the coverage layer and then clicking on the **Data Layer Properties button**. You could also create a Map (in the **Map** tab of the Project Browser) and add the coverage layer to a map to see how it looks.

Creating and running designs

You're now ready to create some designs and run them to calculate coverage probability or to generate single realizations of the design ("surveys"). For examples of this process, see Chapter 3 - Getting Started.

Chapter 7 - Analysis in Distance

Introduction to Analysis in Distance

Distance is designed to promote interactive modeling of distance sampling data. It is easy to set up and run many different models, possibly using different subsets of the data, and to compare and catalogue the results.

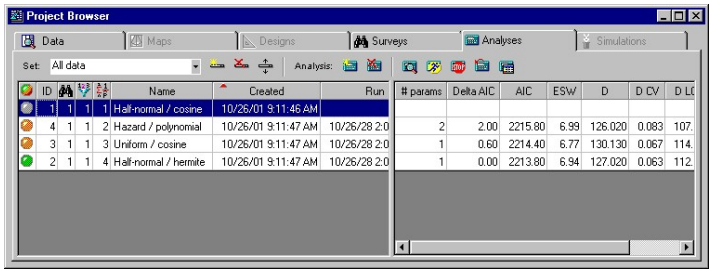
In this chapter, you will find general information about the way that analyses are set up and stored in Distance, and the kind of analyses that you can perform. This chapter is essential reading if you want to get the most out of the program.

More detailed information about each of the analysis windows is in the Appendix - Program Reference.

Introduction to the Analysis Browser

In Distance, your data modeling is divided into individual analyses. Each analysis has a name, and a set of inputs associated with it. If you have run the analysis, it will also have some results.

The main interface for creating, managing and comparing analyses is the **Analysis Browser**, which you can access by clicking on the **Analyses** tab in the **Project Browser**. This displays a summary of each analysis, in table form:


Project Browser
Data
Maps
Designs
Surveys
Analyses
Simulations
Set: All data
Analysis:
ID
Name
Created
Run
params
Delta AIC
AIC
ESW
D
D CV
1 1 1 1 Half-normal / cosine 10/26/01 9:11:46 AM
4 1 1 2 Hazard / polynomial 10/26/01 9:11:47 AM 10/26/28 2:0 2 2.00 2215.80 6.99 126.020 0.083 107
3 1 1 3 Uniform / cosine 10/26/01 9:11:47 AM 10/26/28 2:0 1 0.60 2214.40 6.77 130.130 0.067 114
2 1 1 4 Half-normal / hermite 10/26/01 9:11:47 AM 10/26/28 2:0 1 0.00 2213.80 6.94 127.020 0.063 112

Example of the Analysis Browser, from the Ducknest sample project

In this example, there are four analyses. The first analysis, which is highlighted in blue, has not been run – its status light (left hand column) is grey, and the results columns (right columns) are blank. The next two analyses have been run, but the run generated some warnings – the status light is amber. If an analysis encounters an error during a run, the status light will be red. The last analysis ran with no errors or warnings – its status light is green.

You may also notice in the example that the toolbar along the top of the **Analysis Browser** has a box labelled "Set**:** All data". In Distance you can group your analyses into different Sets. If you were to click on the down arrow beside "All data", you would see that there is another set in this project, called "Truncation at 6 feet". If you chose that set then another table of analyses would

replace those currently displayed. Sets provide a convenient way of grouping related analyses.

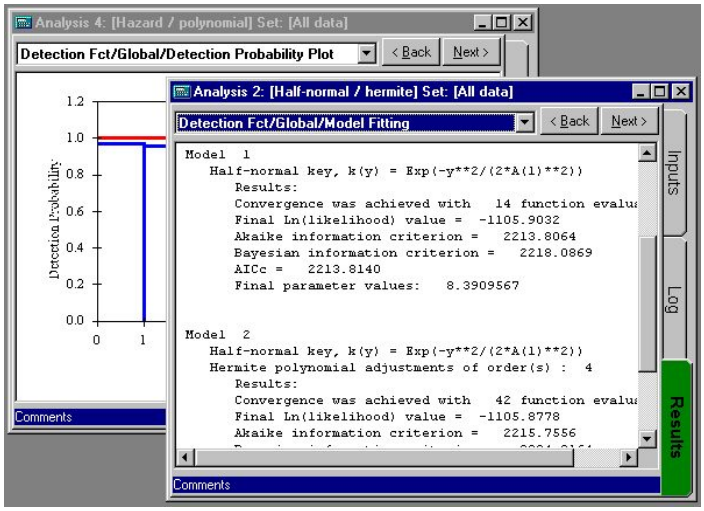
The other buttons in the toolbar allow you to create, delete and manage sets, to create and delete analyses, run analyses, etc.

Introduction to Analysis Details Windows

You can find out more about individual analyses by opening an **Analysis Details** window. You do this either by double-clicking on the status light of an analysis, or by highlighting the analysis you are interested in and clicking the **Show Details** button on the **Analysis Browser**'s toolbar.

Each Analysis Details window contains three tabs (along the right hand side)

- **Inputs Tab**, where you specify how the analysis is to be done.
- **Log Tab**, where you view a log of the analysis once it has been run (useful for pinpointing problems).
- **Results Tab**, where you can read many detailed pages of results from the analysis.



The Analysis Details windows for two analyses, open on different Results pages.

In the above example, the results tab of the top analysis ("Analysis 2") is green because it ran without generating any errors or warnings. For an analysis that has not been run, all three tabs are grey, while if an analysis encounters problems during the run, the Log tab is colored amber (warnings) or red (errors).

Analysis Components

Data Filters, Model Definitions and Survey Objects

In Distance, each analysis is made up of three components: a **Survey**, a **Data Filter**, and a **Model Definition**.

- NEW! **Survey** objects tell Distance what kind of survey you performed (e.g., point transect or line transect), and where the data are stored in the project (which data layers and fields).
- **Data Filters** manipulate the survey data before passing it to the analysis. For example, you could define a Data Filter that discards

all observations with perpendicular distance greater than a given distance (i.e., performs right truncation), or one that selects only data from one survey region.

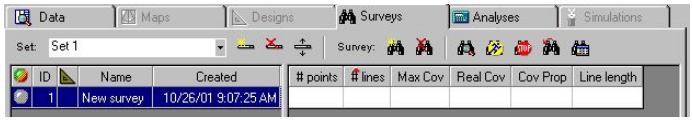
- **Model Definitions** tell Distance how to analyze the data that passes through the Data Filter. Model Definition options include the analysis engine to use, the type of detection function model (e.g., half-normal with cosine adjustments) and the method of estimating variance (analytic vs. bootstrap), as well as many other options.

Each Survey, Data Filter and Model Definition can be attached to one or more analysis (in fact, they can also be attached to no analyses!). Changing the properties of a Survey, Data Filter or Model Definition affects all of the analyses that it is attached to.

Example

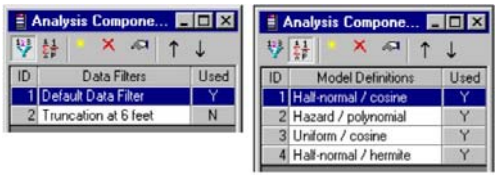
For example, in the Ducknest project (Ducknest.dst, in the Sample Projects folder), one Survey, two Data Filters and four Model Definitions have been defined.

You can get a list of the Surveys by clicking on the **Survey** tab of the **Project Browser**:



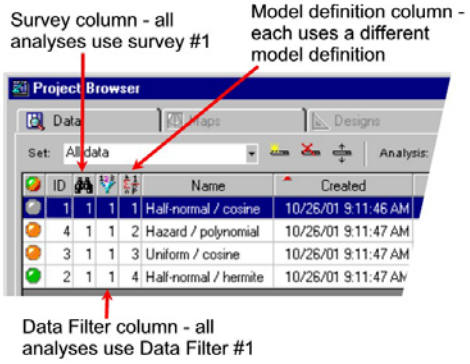
Survey Browser in the Ducknest project.

You can get a list of Data Filters and Model Definitions by clicking on the **View Analysis Components** button on the main toolbar. This opens the **Analysis Components** Window:



Analysis Components window, showing a list of the two Data Filters (left) and four Model Definitions (right), in the Ducknest project.

Using the **Analysis Browser**, we can see that the four analyses in the Analysis Set "All data", all use Survey number 1 (called "New Survey"), and Data Filter number 1 (called "Default data filter"). However, each one uses a different Model Definition.


Survey column - all analyses use survey #1
Model definition column - each uses a different model definition
Project Browser
Data
Maps
Designs
Set: All data
Analysis:
ID
Name
Created
1 1 1 1 Half-normal / cosine 10/26/01 9:11:46 AM
4 1 1 2 Hazard / polynomial 10/26/01 9:11:47 AM
3 1 1 3 Uniform / cosine 10/26/01 9:11:47 AM
2 1 1 4 Half-normal / hermite 10/26/01 9:11:47 AM
Data Filter column - all analyses use Data Filter #1

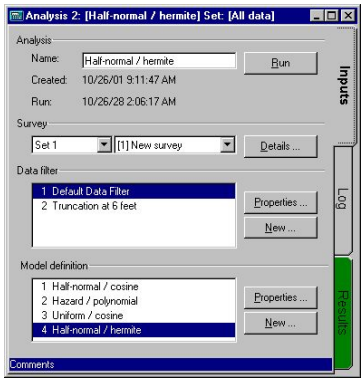
Part of the Analysis Browser from the Ducknest project

If you were to change the Data filter “Default data filter” in some way, then all four of these analyses would be affected (and any other analyses in other Analysis Sets that use this Data Filter). For one thing, the three analyses that have already been run (those with green or orange status lights) would have to be reset and their results deleted, because these results would be out of date. In Distance, if you change a Survey, Data Filter or Model Definition that is being used by analyses that have already been run, Distance automatically issues a warning message and asks you if you want the analyses to be reset.

Working with Data Filters and Model Definitions

Selecting a Data Filter or Model Definition for your Analysis

In Distance, you select the Data Filter and Model Definition for an analysis in the **Inputs** tab of its **Analysis Details** window. For example, the following analysis has the Data Filter called "Default data filter" and the Model Definition called "Half normal / hermite" selected.



Example of the Analysis Details Inputs tab for an analysis, from the Ducknest sample project

The analysis is also called “Half normal / hermite” (top of the picture, in the title bar and beside **Name:**). This is the name that appears in the **Analysis Browser**, so it is always a good idea to give the analysis a name that lets you distinguish it from other analyses in the Analysis Set.

If you wanted to change, say, the Data Filter for this analysis to “Truncation at 6 feet”, you would click on that Data Filter in the list. In the case of the analysis shown above, it would not be a good idea to change the Data Filter as the analysis has already been run and has results associated with it (you can tell this because the Results tab is green). If you change the selected Data Filter or Model Definition in an analysis that has already been run, then Distance will warn you that the results have become out of date and ask whether you want them deleted.

The best way to do an analysis with a new combination of Data Filter or Model Definition is to create a new analysis in Distance for this combination. You do this in the **Analysis Browser**, by clicking on the **New Analysis** button . This automatically creates a new analysis, based on the one that you currently have selected in the **Analysis Browser**. You can then open up the **Analysis Details** for the new analysis. Because the new analysis has not yet been run, you are free to choose the combination of Data Filter and Model Definition you want. If this seems a little confusing, take a few moments to try creating a new analysis in the **Analysis Browser** for the Ducknest project.

Creating New Data Filters and Model Definitions

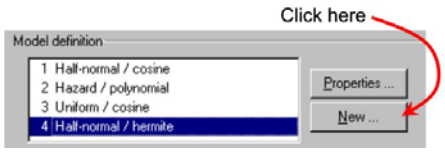
Imagine that in the Ducknest example, above, you have created a new analysis based on the analysis called "Half-normal / hermite", above. Instead of selecting one of the existing Model Definitions for this analysis, you wish to create a new one. Perhaps this new one will use the same detection function model, but will use bootstrapping to estimate the variance of the density estimate.

There are two ways to create a new Model Definition (or Data Filter):

- by clicking the **New...** button on the **Inputs** tab of the **Analysis Details** window of the new analysis. This is described in more detail below.
- by using the **Analysis Components** window. This is described in the section Using the Analysis Components Window, below.

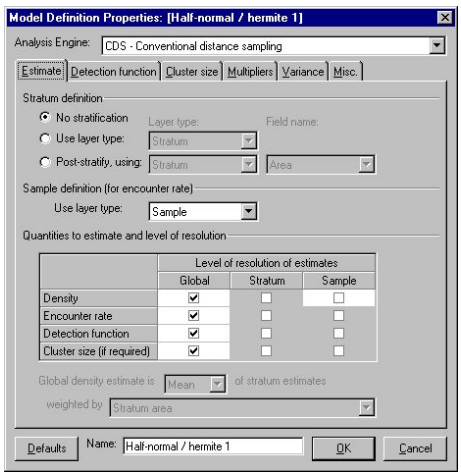
Creating a New Model Definition from the Inputs tab of the Analysis Details window

To create a new **Model Definition**, you click on the **New...** button under **Model definition** on the **Inputs** tab of the **Analysis Details** window:



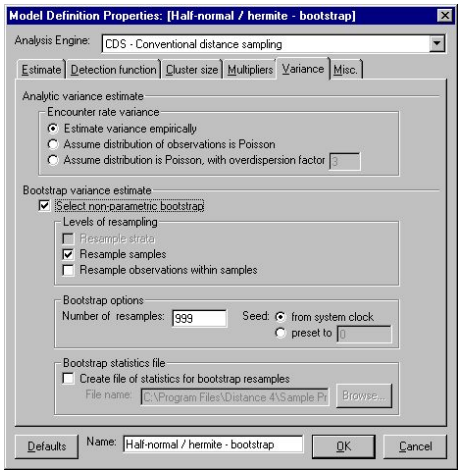
Model Definition section of the Inputs tab on the Analysis Details window

Distance creates a new Model Definition, based on the one you currently have selected, and opens the **Model Definition Properties** dialog for this Model Definition:



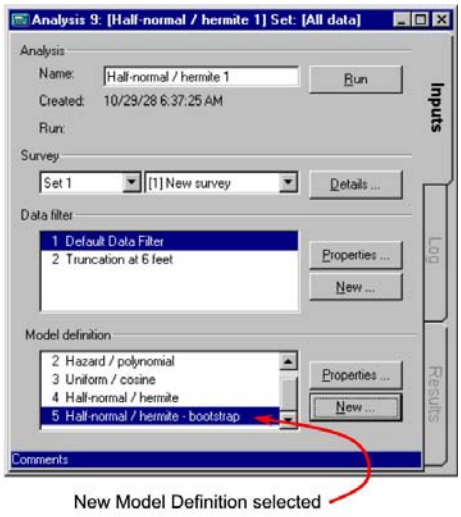
Example Model Definition Properties dialog

You can then edit the properties to reflect the changes you want. As in our example, to change the variance option to use boostrapping, you click on the **Variance** tab and tick the **Select non-parametric bootstrap** checkbox. (More details about the options in this dialog are given in the Model Definition Properties Dialog section of the Program Reference.) You may also want to change the name to reflect the change in properties, for example calling the anlysis "Half-normal / hermite – bootstrap". You do this by editing the **Name:** text box.



Example Model Definition Properties dialog, with bootstrapping option selected

You can now press the **OK** button to save the new options and close the Model Definition Properties dialog. The new Model Definition is automatically selected in the Analysis Details:



Example of the Analysis Details Inputs tab, with a new Model Definition selected

You can then run the analysis by pressing the **Run** button, or you can close the Analysis Details window and run the analysis from the **Analysis Browser** by pressing the button on the **Analysis Browser**'s toolbar.

Creating new Data Filters is exactly analogous to the process just described for Model Definitions. In this case, when you press the Data filter **New...** button, a new Data Filter is created based on the one currently selected, and the Data Filter Properties dialog opens. To find out more about the Data Filter options, see the Data Filter Properties Dialog section of the Program Reference.

Viewing and Editing Existing Data Filters and Model Definitions

In some cases, you may want to view or edit the properties of an existing Data Filter or Model Definition. For example, you may wish to check the options in a Model Definition before selecting it for an analysis. There are two ways to do this:

- by clicking the **Properties...** button on the **Inputs** tab of the **Analysis Details** window. This will open the properties dialog

of the Data Filter or Model Definition that is currently selected for that analysis.

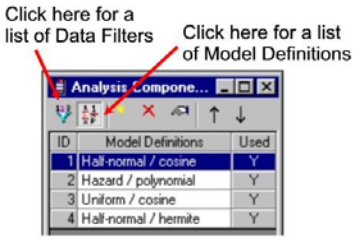
- by using the **Analysis Components** window. This is described in the section Using the Analysis Components Window, below.

Note! If you change the options in a Data Filter or Model Definition that is being used by any analyses that have been run, Distance will warn you when you press the OK button that the analyses will be reset.

To find out more about the options available in the **Data Filter Properties** and **Model Definition Properties** dialogs, see the Program Reference sections Data Filter Properties Dialog and Model Definition Properties Dialog.

Using the Analysis Components Window

The **Analysis Components** window is designed to be a convenient way of manipulating Data Filters and Model Definitions. The window shows a list of all Data Filters or Model Definitions in your project. Using buttons on the toolbar you can create, delete, view and arrange the listed components.


Click here for a list of Data Filters
Click here for a list of Model Definitions
Analysis Compone...
ID
Model Definitions
Used
1 Half-normal / cosine Y
2 Hazard / polynomial Y
3 Uniform / cosine Y
4 Half-normal / hermite Y

Analysis Components window, showing a list of the Model Definitions in the Ducknest sample project

In the **Analysis Components** toolbar, you click on the first button to get a list of all the Data Filters, and the second button to get a list of all the Model Definitions in your projects. The other buttons allow you to copy (i.e. create), delete, view and arrange (move up and down the list) the component that you have selected.

You can also work with Data Filters and Model Definitions in the **Analysis Details** window. Using the **Analysis Components** window is most useful when you have a large number of components in your project, as you can arrange them into a logical order, delete the ones you are not using, and easily rename them.

Tip! The last column in the table of analysis contents tells you whether that component is currently being used in any analyses: "Y" means it is being used and "N" means that it is not. This is useful because when there are many components (e.g., many Model Definitions if you have been doing a lot of analyses), it is easy to loose track of which are being used and which are no longer required. Also, if you double-click on a "Y", you get a list of the analyses that use that component.

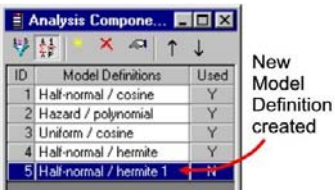
Example

In the section Creating New data Filtes and Model Definitions we showed how to create a new Model Definition and associate it with a new Analysis using the

Inputs tab of the **Analysis Details** window. Here, we show how to do the same thing using the **Analysis Components** window.

In the Ducknest sample project, there are four analyses in the set "All data". These analyses use four different Model Definitions, each with a different detection function model. Imagine you want to add a new analysis, with a new Model Definition. Your new analysis will be based on the "Half-normal / hermite" analysis, but will use bootstrapping to estimate the variance of the density estimate.

You begin by creating a new Model Definition. In the Analysis Components window, you highlight the Model Definition "Half-normal / Hermite" and click on the **New Item** button . A new Model Definition is created, based on the one you had highlighted:



Analysis Components window, showing the new Model Definition

The new Model Definition has been given the default name "Half-normal / hermite 1". You may want to change the name to reflect the options you're about to set – double click on the name and type "Half-normal / hermite – bootstrap".

You can now edit the new Model Definition properties, by double-clicking on the ID of the new Model Definition, or by clicking the **View Item Properties** button . The **Model Definition Properties** dialog opens. To change the variance option to use boostrapping, you click on the **Variance** tab and select the "non-parametric bootstrap" option. (More details about the options in this dialog are given in the Model Definition Properties Dialog section of the Program Reference.) You can now press the **OK** button to save the new options and close the dialog.

You have now set up the new Model Definition ready for use. If you want to perform other types of analysis, you could set up more Model Definitions at this point.

You now need to create a new Analysis, and attach the new Model Definition to this new Analysis. In the Analysis Browser (i.e., the **Analysis** tab of the Project Browser), click on the **New Analysis** button . Double-click on the status button of the new analysis – this opens the Analysis Details window. In the **Model definition** section, select your new model definition, and in the **Name:** section, type in a suitable name for your new analysis, e.g., "Half-normal / hermite – bootstrap". You can now run the analysis.

This approach to setting up new Model Definitions (or Data Filters) is most useful when you have several to set up at once. You can use the **Analysis Components** window to set up your new components, then use the **Analysis Browser** to create new Analyses, and associate the new analyses with the new components. Then, in the **Analysis Browser**, you can highlight all the new Analyses, click the run button , and go and have a cup of tea while they all run! (For more about running analyses, see Running Analyses on page 12).

Working with Surveys during Analysis

Surveys tell Distance what kind of survey you performed (e.g., point transect or line transect), and where the data are stored in the project (which data layers and fields). Each analysis is associated with a Survey. Surveys are also used when designing new field surveys – for more on this see Chapter 6 - Survey Design in Distance.

In routine analysis of Distance sampling data, you don't have to worry too much about Surveys. If, in the Setup Project Wizard, you tell Distance that you want to analyze a survey that has been completed, Distance will automatically create a Survey for you, based on what you tell it about your survey in the wizard. This Survey will be used by default for all the Analyses you create.

You only need to work with Survey during analysis if:

- you want to set up the project manually. In this case, you will need to create the Survey and set its properties yourself.
- you wish to do analyses involving more than one Survey. Situations where this may be useful are covered in the next section (Analysis with Multiple Surveys).

You can get more information about the Survey associated with your analysis by clicking the **Details...** button on the **Inputs** tab of the **Analysis Browser**. This opens the **Survey Details** window. Click on the **Properties...** button of the **Survey Details** window.

Analysis with Multiple Surveys

Normally, for analysis in Distance, you only need one Survey. This Survey tells Distance what type of survey you performed and where the data from the survey are stored. However, there are some situations when it is useful to have more than one Survey in a project. Examples include:

- you have a complicated data structure, for example with two or more layers of type Sample (e.g., for two or more survey regions or years). In this case, you will set up one Survey to point to each sample layer. You could then create one Analysis set for analyses that use the first layer, and another for analyses that point to the second.

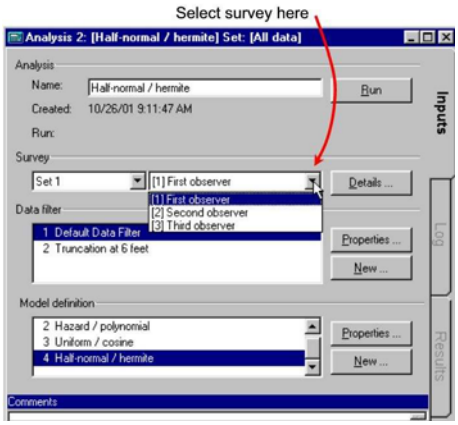
Having multi-site or multi-year data in separate data layers means you will not be able to do a single combined analysis of all data. For this reason, it is often better to put this type of data into a single data layer, with an extra field indexing the year or site number. You can then use the data selection page of the Data Filter to select out individual years/sites or combinations of years/sites for analysis as required.

- you have multiple measurements of the same objects, for example from multiple observers. One case where this can occur is in field trials on artificial objects, where multiple observers traverse the same lines or points. Your data will then have multiple "distance" fields. You can set up a Survey for each distance field (in the **Data Fields** tab of the Survey), and then analyze each observer separately.
- you have clustered data, but you want to ignore the clustering for some analyses, and just calculate density of clusters. In this case, you would create a new survey with the Observations option set to Single objects.
- you have multi-species data where some species occur in mixed species groups (see Zero Cluster Sizes in CDS Analysis for more on this).

If you do want to set up multiple Surveys, you do this from the **Survey Browser**, which you access by clicking on the **Survey** tab of the **Project Browser**.

- To create a new survey, click on the **New Survey** button.
- To edit the properties of a survey, click on the **Show Details** button in the **Survey Browser**. A **Survey Details** window opens. On the **Input** tab, click on the **Properties...** button.

Once you have created and set up the Surveys you require, you associate them with Analyses by selecting the appropriate Survey from the list on the **Survey** section of the **Analysis Details** window.

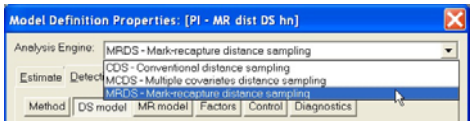


Example of Analysis Details Inputs tab, showing selection of Survey object

Analysis Engines

In Distance, you have a choice of **analysis engines** to perform an analysis. Each analysis engine has different capabilities, and has different inputs and outputs. Distance has three analysis engines built in: a conventional distance sampling (CDS) engine, a multiple covariate distance sampling (MCDS) engine, and a mark recapture distance sampling (MRDS) engine. More engines are planned for future versions of Distance.

You choose the analysis engine when you are setting up a model definition – select the appropriate engine from the drop down list at the top of the Model Definition Properties dialog:



Choosing an analysis engine

A description of the options available for each engine is given in the Model Definition Properties Dialog section of the Program Reference.

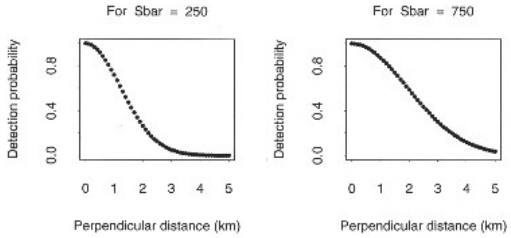
Conventional Distance Sampling (CDS) Engine

This engine provides a design-based analysis of line or point transect data, using the approach described by Buckland et al. (1993, 2001). Probability of detection is modeled as a function of observed distances from the line or point, using robust, semi-parametric methods. One level of stratification is allowed, and there are various methods for dealing with cluster size bias. Variance can be estimated empirically, or via a non-parametric bootstrap.

For more information about this engine, see Chapter 8 - Conventional Distance Sampling Analysis.

Multiple Covariate Distance Sampling (MCDS) Engine

This engine contains (almost) all the features of the CDS engine, but also allows additional covariates to be included in the detection function model, in addition to observed distance. These covariates enter through the scale parameter of the key function (via a log link function). This means that the covariates are assumed to influence the *scale* of the detection function, but not its *shape* (see picture, below).



Example estimated detection functions, where cluster size (Sbar) is the covariate. The basic shape of the function is the same (half-normal), but the effective strip width is wider at cluster size 750.

For more information about this engine, and when it can be useful, see Chapter 9 - Multiple Covariates Distance Sampling Analysis. This engine was first introduced in Distance 4.0.

Mark Recapture Distance Sampling (MRDS) Engine

This engine permits analysis of data collected from two survey platforms, where the assumption of certain detection detection of objects on the trackline can be relaxed. CDS and MCDS analyses are also possible, although adjustment terms are not currently available to modify the shape of the detection functions in this engine.

For more information about this engine, see Chapter 10 - Mark Recapture Distance Sampling. This engine was first introduced in Distance 5.0.

Density Surface Modelling (DSM) Engine

This engine allows users to model variation in density in relation to covariates, using distance sampling (and plot sampling) data.

For more information about this engine, see Chapter 11 - Density Surface Modelling. This engine was first introduced in Distance 6.0.

Running Analyses

Once you have created an analysis, and set it up by choosing an appropriate Survey, Data Filter and Model Definition, you are then ready to run it.

There are two ways to run an analysis in Distance:

- in the **Analysis Browser**, select the analysis and press the **Run Analysis** button (or choose **Analyses** | **Run Analysis**)
- in the **Inputs** tab of the **Analysis Details** window of that analysis, press the **Run** button **Analysis - Inputs** | **Run Analysis**.

Tip! You can run more than one analysis at once. Simply highlight all of the analyses you want to run in the Analysis Browser and then press the Run Analysis button. This is useful if you are planning on doing a number of long analyses - simply set them all up, select and run them all at once, and go and have a cup of tea!

Aside! NEW! In Distance, when you select more than one analysis to run at once, one is run straight away, and the others are queued up to run in turn. You will see the status light of the queued analyses turn to a "q". This change was made for two reasons. Firstly, it makes more efficient use of a single-processor computer to have only one computer-intensive process running at a time. Secondly, Distance is designed to have numerical engines running that require exclusive access to the distance database. One example of this is the survey design engine. Because of this, only one item (design, survey, analysis) can run at a time.

Note! In Distance, Analyses run in a background process. This means that you can carry on working with the interface while the analysis is running. The only time you should go easy on the interface is when an analysis is initializing or finishing (you can tell when this is happening because the pointer is turned into an hourglass and a message is printed on the status bar at the top of the screen). This feature is particularly useful if you are doing analyses such as bootstraps that take a long time to run.

Tip! Windows NT, 2000 and XP are generally better built than Windows 95, 98 and ME. However, one unfortunate consequence of this for us is that background processes tend to get more CPU time. This means that the Distance interface may slow down quite noticeably while an analysis is running under Windows NT/2000/XP, even on well-configured machines. You can give the interface a boost by changing the Foreground performance setting in the Performance tab of the Windows System Properties dialog (System icon in Control Panel) to Maximum.

Tip! You can stop an analysis that is running by pressing the **Stop** button in the **Analysis Details** window (this button replaces the **Run** button while the analysis is running), or by highlighting the analysis in the **Analysis Browser** and pressing the **Reset Analysis** button. However, on some systems, the analysis will appear to stop but will carry on running in the background, using up system resources. For more about this, see Stopping an Analysis, in Chapter 10.

Locking the Data Sheet

A complete analysis of distance sampling data in Distance usually proceeds in three stages:

- creating and setting up the distance project file

- getting the data into Distance
- analyzing the data

Clearly, it is generally not a good idea to change the data after it has been analyzed. Because of this, Distance has built-in safeguards to prevent you from accidentally editing the data once the analysis phase has begun. By default, whenever you perform an analysis, Distance locks the Data Sheet. This means that the data cannot be edited or deleted. You can tell that the data have been locked, as the Lock Data Sheet button in the Data Explorer toolbar becomes depressed.

In some circumstances you may want to change the data after the analysis phase has begun. For example, you may discover an error in the data entry as a result of some exploratory analysis. You may also want to add another year's data to your project. You may want to add additional columns, in order to perform post-stratification. If you wish to change the data, it is simple to unlock the data sheet: click on the depressed button to pop it back out.

If you regularly update the data during the analysis stage then you can disable the automatic locking feature of Distance, by unchecking the option **Automatically lock data sheet whenever an analysis is run** in the Analysis tab of the Preferences dialog (choose **Tools** | **Preferences** from the main menu bar).

Cleaning the Temp folder

When analyses are run, temporary files are written to the Windows temp folder (the location of this folder is system-dependent). As analyses finish, these temporary files are deleteted. If an analysis crashes, then the temporary files may not be deleted correctly, so over time they may build up in the temporary folder. To clear all temporary Distance files from this folder, choose **Tools** | **Clean temp folder ...**.

R Statistical Software

Distance has a link to the free statistical software R. The Mark Recapture Distance Sampling (MRDS) analysis engine is implemented as an R library, and so you must have a working copy of R installed on your computer before you can use that engine.

You can download and install R after you have installed Distance, and you can run Distance without having R installed - but you will get an error message if you try to run the MRDS engine without R.

R is under very active development, and new versions are released quite frequently. Unfortunately, new versions are sometimes not compatible with libraries compiled in old versions. We will endeavour to test our libraries with each new version as it appears, and update it as required. For more information about the version we are currently supporting, please browse to the Program Distance Web Site, Support, Updates and Extras page.

To use the MRDS engine, you don't have to know anything about R beyond how to install it. However, R is a fully featured, widely-used statistics package which you may consider using for your other analyses. You can find out more about R from the R project home page, http://www.r-project.org/.

Installing and Configuring R

Instructions for installing R are given in the file readme.rtf in the Distance program folder. You can also access this file from within Distance by choosing **Help** | **Release notes**.

Before installing, you should check which versions of R are currently supported by Distance. Latest information is on the Support, Updates and Extras page of the Program Distance Web Site.

Once you have R installed, you should be able to run MRDS analyses straight away from within Distance. Distance automatically recognizes the latest version of R you have on your system the first time you run an MRDS analysis or open the Preferences dialog, and automatically installs the mrds library. Once it has registered the presence of R, it keeps using the same version until you manually tell it to update (see Updating the Version of R That Distance Uses). You can check which version it is using by choosing **Tools** | **Preferences** and looking under the **Analysis** tab.

If you cannot see the plots that R produces, see Images Produced by R

Updating the Version of R That Distance Uses

If you install a new version of R after running some analyses using an older version, Distance does not automatically switch to the new version. This is because Distance might not be compatible with the newer version of R. Before telling Distance about the new version of R, you should check it is compatible, by looking on the Support, Updates and Extras page of the Program Distance Web Site. Having checked, you can make Distance use the new version by choosing **Tools** | **Preferences**, choosing the **Analysis** tab and updating the **Folder containing R** to the folder that contains the new version. Distance will automatically install any required libraries the first time R is run from within Distance. Once you have checked the new version of R works, you can delete the folder containing the old version.

Contents of the R folder

Distance projects that contain analyses run with R (e.g., MRDS analyses) have a folder within the project data folder called "R". This folder contains:

- the R object file, .RData. This file holds all the objects created by R for that project. By default, new objects that are created for an analysis are deleted at the end of that analysis, so the .RData file is virtually empty. However, this default behaviour can be changed - see Analysis Preferences Tab.
- image files generated by R. These files are loaded into the Results tab when the Analysis Details for an analysis is opened. For more about these files, see Images Produced by R.

Images produced by R

The images produced by R are stored as files in the R folder (see Contents of the R folder). They are of the general form

```
[prefix].[analysis ID].[plot number].[suffix]
```

for example qq plot 1 for analysis 8 in windows metafile (.wmf) format would be qq.8.01.wmf.

These files are loaded into the Results tab when the Analysis Details for an analysis is opened inside Distance.

The image files can be used in producing manuscripts and other reports of analyses done. Many aspects of the images can be changed - see below.

Changing the image properties

By default, R produces images in Windows Metafile format (.wmf). This is a vector format that can be viewed at a variety of sizes without loss of quality. However, .wmf files do not display properly on older Windows operating systems, so you may wish to switch to another format. You may also wish to switch to a more compact format to save disk space.

In addition, many other aspects of the images can be configured, such as the image size (this only affects the size in the image file; images are automatically scaled to fill the Results tab of the Analysis Details window), line width, font size, etc.

To change image properties, choose **Tools** | **Preferences**, **Analysis** tab, and under **R Software** click on **Image Properties…**. For more about the options, see R Image Properties Dialog in the Program Reference.

Chapter 8 - Conventional Distance Sampling Analysis

Introduction to CDS Analysis

Conventional distance sampling (CDS) analysis refers to analysis of distance sampling data using the methods described by Buckland et al. (1993, 2001). Probability of detection is modeled as a function of observed distances from the line or point, using robust, semi-parametric methods. The distances can be recorded exactly, or grouped into non-overlapping intervals (also called "bins"). Various methods are described for dealing with data where the objects are clusters, rather than individuals.

The CDS analysis engine in Distance implements these methods. Detection function, encounter rate and cluster size (where relevant) are estimated separately, and the results are combined to estimate density or abundance. Additional factors that influence density estimation (such as violation of the assumption that all object at the point or line are seen) can be incorporated through the use of multipliers.

The CDS engine allows one level of stratification – strata may be, for example, geographic regions. Detection function, encounter rate and/or cluster size can be calculated either by stratum or globally, and density can be estimated globally, by stratum and/or by sample (a sample in this context is an individual line or point). Variance can be estimated analytically, or via a non-parametric bootstrap. The bootstrap can also be used to produce point and interval estimates that include model selection uncertainty (see Model averaging in CDS Analysis).

The mechanics of setting up and running analyses in Distance is outlined in Chapter 7 - Analysis in Distance. This chapter covers topics that are specific to the CDS engine, such as guidelines for approaching CDS analysis, how to deal with grouped (binned) data, stratification and multipliers. (Many of these topics also apply to the multiple covariates MCDS engine, so we recommend reading this chapter as well as the next one before embarking on any MCDS analyses.)

This manual is designed to complement the standard text on Distance sampling (Buckland et al. 1993 or 2001). Users of Distance are referred to that text for a detailed explanation of conventional distance sampling, and an extensive set of examples.

The CDS and MCDS engines are implemented as a single FORTRAN program which is run from within the Distance interface. You can also run the engine as a stand-alone program – for details see the Appendix - MCDS Engine Reference.

Aside! It is also possible to perform a CDS (or MCDS) analysis using the MRDS engine – see Single Observer Configuration in the MRDS Engine in Chapter 10 of the Users Guide.

Modelling the Detection Function

A central part of the analysis of distance sampling data is modeling of the detection function. The CDS engine implements the robust key function + series expansion (adjustment term) approach outlined by Buckland et al. (1993, 2001).

The candidate key functions offered are Uniform, Half-normal, Hazard-rate and Negative exponential (this last function is not recommended, except for salvage analyses; see Buckland et al. 1993, 2001). The candidate series expansions are Cosine, Simple polynomial and Hermite polynomial. The user is free to choose any combination of key function and series expansion, and there are a wide range of options for both automatic and manual selection of the number and order of series expansion terms that are fit to the data. However it is not necessary or desirable to try every possible combination! A list of suitable candidate models is presented in Buckland et al. (1993, 2001), and their selection illustrated throughout the book.

In Distance, you implement a detection function model by defining a Model Definition and then selecting the appropriate options from the **Detection Function Model** tab. These options are outlined in the Program Reference section on the Model Definition Properties Dialog.

Setting up a Project for CDS Analysis

For an example of how to set up a project for CDS, see Getting Started Example 1: Using Distance to Analyze Simple Data in Chapter 3. You should also read Chapter 4 - Distance Projects, Chapter 5 - Data in Distance and Chapter 7 - Analysis in Distance.

CDS Analysis Guidelines

The following is a condensed version of the guidelines discussed in Section 2.5 of Buckland et al. (2001), with some specific recommendations regarding the organization of analyses in Distance. We have steered clear of defining specific "cookbook" procedures for data analysis and, above all, recommend that you *do not unthinkingly use the Distance defaults*.

Generally, we recommend that you start by thoroughly exploring your data, by plotting histograms of recorded distances with the data sub-divided into a large number of intervals. This can be done in Distance by creating a **Model Definition** with an arbitrary model and manually defining a large number of intervals in the **Detection Function**, **Model Diagnostics** page of the **Model Definition Properties** dialog. Look for evidence of heaping, evasive movement, outliers and possible gross errors. Line transect data that are recorded as perpendicular distance and angle are better examined outside of Distance, where the distribution of angles can be inspected and problems such as rounding to zero degrees can be diagnosed.

At this stage, you should not be concerned with estimating density – indeed any estimates produced are often distracting and misleading. Therefore, it is good practice to tell Distance not to estimate density, by un-checking the boxes in the **Density** row of the section labelled **Quantities to estimate and levels of resolution** on the **Estimate** tab.

The issue of suitable truncation can be examined by creating **Data Filters** with different truncation distances. Similarly, it may prove beneficial to group exact distance data into intervals - this is also done in the Data Filter. This exploratory phase is open ended, but you should strive to fully understand the data and possible violations of the assumptions of distance sampling analyses (Buckland et al. 2001, Section 2.1). In Distance, it may be worth grouping these exploratory analyses into a suitably named Analysis Set.

Once the data have been properly prepared, and a decision has been made about truncation and other Data Filter issues, the model selection phase can begin. We recommend selecting a small number of sensible candidate models from those available in Distance, and defining a separate Model Definition for each one. This way, a separate analysis can be created for each model, and the AIC and Delta AIC (or AICc and Delta AICc) columns in the **Analysis Browser** can be used to sort and compare the analyses. Of course, other criteria should also be used in selecting among the candidate models, such as goodness of fit (especially near zero distance). A great deal of useful information about each model is stored in the **Analysis Details**, **Results** tab.

In many cases these analyses will suggest additional explanatory work, so the process of model selection and exploration is often iterative. Other issues, such as the appropriate levels for estimating parameters (sample, stratum, global) must also be considered (see Stratification and Post-stratification in this Chapter for a discussion of some of these issues).

As the number of analyses defined and run starts to build up, it becomes worth considering grouping the analyses into different Analysis Sets in the **Analysis Browser**. The **Analysis Components** window can also be used to move related Data Filters and Model Definitions so that they are positioned adjacent to one another. The **Comments** section of the **Analysis Details** window for each Analysis can be used to record pertinent information, such as what you learnt by running the analysis.

At some point, you select a model you believe to be the best for the data set under consideration. This is the time to consider making bootstrap estimates of variance (see **Model Definition**, **Variance** Tab - CDS and MCDS in the Program Reference), and beginning to make inferences from the abundance estimates produced. In many cases there will be perhaps two or three models that appear to fit the data equally well. Distance allows you to define multiple models in the **Model Definition**, **Detection Function Models** tab - if you create a Model Definition that includes all of the final models and specify bootstrap variance estimates, then the estimated variance will account for this uncertainty in model selection as well as the other sources of variation. This approach has much to recommend it. (For more on this, see Model Averaging in CDS Analysis in this Chapter).

The above guidelines give a broad overview of how the analyst might proceed. These ideas are developed much more fully in Buckland et al. (1993, 2001), and extensive examples are given to illustrate the approach.

Output from CDS Analyses

The CDS engine produces the following output:

- a summary of results in the **Analysis Browser**. For general information about the Analysis Browser, see the section Introduction to the Analysis Browser in Chapter 7. There are many results statistics available, and you can select which ones are shown independently for each analysis set using the Column Manager (see Column Manager Dialog in the Program Reference). An explanation of some of the columns is given in the section CDS Analysis Browser Results.

- a detailed listing of results in the **Results** tab of the **Analysis Details** window. These are described in the following section, CDS Results Details Listing.
- a log of the analysis, highlighting any possible problems, in the **Log** tab of the **Analysis Details** window. For information about troubleshooting problems, see Chapter 12 - Troubleshooting.
- (optionally) text files, containing the results listing, analysis log, summary statistics, bootstrap statistics and plot data. For more about these, see the section on Exporting CDS Results.

CDS Results Details Listing

When an analysis has run, a great deal of information is available in the **Results** tab of the **Analysis Details** window. This information is split into pages, as follows:

- **Estimation options listing.** Gives a summary of the analysis options you chose.
- **Detection Fct.** A set of detection function pages for each subset of data used for modeling the detection function. If the detection function is estimated globally, there are one set of these pages. If by stratum, there is one set for each stratum.
  - **Model fitting.** Gives details of the models fit, and the final model selected.
  - **Parameter estimates.** A summary of the parameter estimates for the final model selected, including correlations among estimates.
  - **Plot: Qq-plot.** (Not for interval data) Qq-plots are another graphical method for assessing model fit – for more information, see CDS Qq-plots in this Chapter.
  - **K-S GOF Test**. (Not for interval data) Kolmogorov-Smirnov and Cramér-von Mises tests of goodness-of-fit. For more information, see CDS Goodness of fit tests in this Chapter.
  - **Plot: Detection probability.** Plot of the detection function, superimposed on histograms showing the frequency of counts. (These frequencies are scaled for point transects, see Buckland et al. 2001)
  - **Plot: Pdf.** Probability density function plot – only for point transect data.
  - **Chi-sq GOF test.** Table of observed and expected frequencies in each histogram bin, together with a χ^2 goodness-of-fit test. This test gives a measure of how well the model fit the data, based on a comparison of the observed and expected frequencies of observations within distance bins.

The χ^2 test is known to be biased if expected cell counts are small. If the expected counts are less than 2.0, Distance produces a second table, where adjacent bins are pooled until the expected counts are greater than 2. This procedure is rudimentary, and users can probably construct a better test by hand.

Note! The previous three pages are designed to help you diagnose model fit. By default, you get three sets of these pages, with the data divided into equally spaced intervals, and the number of intervals being $n^{0.5}$, $2/3n^{0.5}$ and $3/2n^{0.5}$ (where n is the count of objects). Instead of using the defaults, we recommend you always define your own cutpoints.

Tip! These pages are probably the most important of the whole results output, and you should check them carefully. Check the Model Fitting output to see if any of the models that were fit did not converge, or hit any of the constraints. Even if the model affected is not the one that was eventually selected (if you're using automatic model selection), the selection process can be affected if there was a problem in the fitting. Check the plot(s) and GOF tables to look for evidence of lack-of-fit, and possible problems with the data such as rounding and evasive movement. These issues are mentioned in this Chapter on the page entitled CDS Analysis Guidelines, and are covered in more detail in the Distance Book.

Tip! If there is a problem in the fitting routine, such as non-convergence, it may be useful to look at the parameter estimates for each iteration of the fitting algorithm. To do this, check the option "Report results for each iteration of the detection function fitting routine" in the Model Definition Misc. Tab, and then re-run the analysis.

Tip! If any of the parameter estimates hit the default upper or lower bounds, you should consider setting bounds manually. You do this in the Constraints page of the Detection Function tab in Model Definitions.

- **Cluster size**. A set of pages for each subset of data used in estimating expected cluster size (e.g., one for each stratum). If objects are not in clusters, these pages are omitted.
  - **Estimates.** Gives the estimated expected cluster size, including the size-bias regression results, if requested (the default).
  - **Regression plot**. Text-based plot showing size-bias regression, if requested.
- **Density estimates.** One per subset of data for which density is estimated (e.g., one for each stratum)
- **Estimation summary.** Set of pages containing tables summarizing the results, and giving variance estimates and confidence limits.
  - Encounter rates
  - Detection probability
  - Expected cluster size (if objects in clusters)
  - Density & Abundance
- **Bootstrap summary.** (Only if variance by bootstrap option selected.) A set of pages similar to the estimation summary, but with bootstrap point estimates, standard errors and confidence limits. The bootstrap point estimate is the mean of the point estimates from the bootstrap replicates (useful for model averaging

– see Model Averaging in CDS Analysis). Two types of confidence limits are given. The first use the bootstrap standard error to generate parametric, log-normal confidence limits. The second use the percentile method – i.e., for x% confidence intervals the (x/2)th and (100-x/2)th quantiles of the bootstrap estimates are given. In general, the latter confidence intervals are considered more reliable.

Aside! The parametric bootstrap confidence intervals on density and abundance use the same degrees of freedom as the original analysis, rather than re-calculating the degrees of freedom using formula 3.75 of Buckland et al. (2001).

Tip! For information about how to export the results text or plots into another program, see Exporting CDS Results from Analysis Details Results.

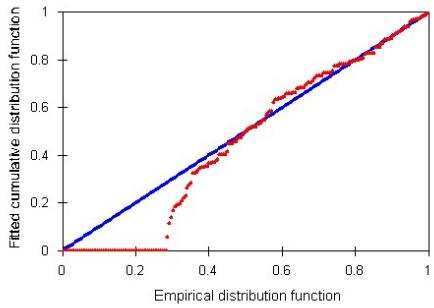
CDS Qq-plots

Distance gives quantile-quantile (qq)-plots for all analyses that use exact data (i.e., those where the data is not transformed into intervals in the Data Filter). Qq plots are useful for diagnosing problems in the data such as rounding to preferred values and other systematic departures from the fitted model. A major advantage of these plots over the histograms of the detection function and probability density function (pdf) is that they do not require the data to be grouped into intervals. A disadvantage is that they require a little effort to understand the output.

In statistics, qq-plots are used to compare the distribution of two variables – if they follow the same distribution, then a plot of the quantiles of the first variable against the quantiles of the second should follow a straight line.

To compare the fit of a detection function model to the data, a standard method is to plot the fitted cumulative distribution function (cdf) against the empirical distribution function (edf). The cdf, F(x), gives the probability of getting a distance less than or equal to x for a given model. The edf, S(x), gives the proportion of the data with distances less than x. (Note – this explanation ignores tied values.) If the data fit the model, then the fitted cdf and edf should be the same.

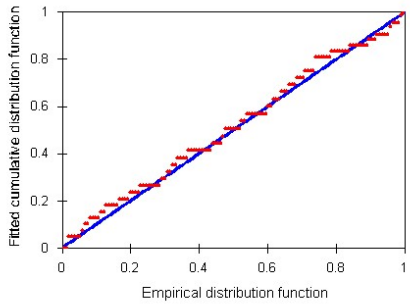
To make the qq-plot, the fitted cdf is evaluated for each observation. The data are then sorted into increasing order i=1, … n and the edf is calculated as (i-0.5)/n. The following plot shows an example where 58 of the 204 data points are at 0 distance (yes, this is a real dataset!). The red dots show the data, and the blue line is where they should lie if the fit of the model was perfect.



Example qq plot, showing a severe 'spike' at 0 distance

The cdf at 0 distance is 0, so the 58 points appear along the bottom left side of the plot. Clearly in this case, the data do not fit the model. This can be confirmed in Distance using the Kolmogorov-Smirnov and Cramér-von Mises goodness of fit tests on the next page of results output.

The following plot shows an example where the fit appears quite good, with most points close to the line and little systematic departure. The data have clearly been rounded (e.g., to the nearest meter) as there are several data points at each level of the cdf. Such rounding should not affect the reliability of the parameter estimates at all.



For more on qq-plots, see Chapter 11 of Buckland *et al.* (2004). For options associated with qq-plots in the CDS engine, see the Program Reference page on Model Definition Properties, Diagnostics - Detection Function Tab - CDS and MCDS. For information about how to export qq-plots (and other output) from Distance into Word processors, spreadsheet and other graphing programs, see Exporting CDS Results from Analysis Details Results.

CDS Goodness of fit tests

Previous versions of distance allowed users to test the fit of a model using χ^2 goodness-of-fit tests. A disadvantage of this test for ungrouped data is that the data must first be put into intervals before the test can be performed, and the selection of cutpoints can have a strong influence on the outcome of the test. Therefore, we have added three tests of goodness-of-fit that operate directly on the exact distances observed. These tests are not produced when the data are analyzed in intervals.

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (k-s) statistic focusses on the largest difference between the cumulative distribution function (the cdf, $F(x)$), and the empirical distribution function (the edf, $S(x)$ – see the previous topic for more on these functions).

To calculate the statistic, the fitted cdf is evaluated for each observation, and these cdfs are arranged in ascending order and indexed i=1, … n. The k-s statistic is then given by

$$\hat{D}_n = \max\left(\hat{D}_n^+, \hat{D}_n^-\right)$$

where $\hat{D}_n^+ = \max\left(i/n - \hat{F}(x_i)\right)$ and $\hat{D}_n^- = \max\left(\hat{F}(x_i) - (i-1)/n\right)$.

The k-s statistic can be used to test whether there is a significant departure between the edf and cdf – in other words whether the data fit the model.

Aside! The upper tail probabilities ("p-values") reported by Distance are calculated by evaluating the following expansion:

$$P\left(D_n > \frac{z}{\sqrt{n}}\right) = 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}$$

where $z = \hat{D}_n\sqrt{n}$ and n is the number of observations (Gibbons 1971, page 81). This should be suitably accurate for practical application for sample sizes of about 35 or greater.

Cramér-von Mises test with uniform weighting function

Unlike the k-s test, which focuses on the largest difference between the cdf and edf, Cramér-von Mises (C-vM) family tests focus on the sum of squared differences between cdf and edf. They are of the form:

$$Q = n\int_{-\infty}^{\infty}[F(x)-S(x)]^2\psi(x)dF(x)$$

where $\psi(x)$ is a weighting function that allows us to give different weights to different parts of the distribution. If we give all observations the same weight then we obtain the standard C-vM statistic, calculated by:

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[\hat{F}(x_i) - \frac{i-0.5}{n}\right]^2$$

Aside! The tail probabilities for W^2 depend on sample size, and a closed form expression for them is not available. Instead, we used simulation to estimate critical values for W^2 at a range of sample sizes from 5 to 1000 and at α-levels of 0.001, 0.005, 0.01, 0.025, 0.05, 0.1, 0.15, and 0.2, 0.3, … , 0.9. These values are stored in Distance, and the program uses linear interpolation to construct a set of critical values for the observed sample size. It then reports which of the critical values the observed W^2 falls between, using the form: "α_l < p <= α_h" where α_l and α_h are the bounding critical values.

Cramér-von Mises test with cosine weighting function

This is similar to the above test, but with a weighting function that puts more emphasis on observations closer to 0 distance. The rationale is that these observations have more influence on the estimated value of f(0) or h(0), so we expect this test to have more power to detect influential departures from the fitted function, and be more robust to departures at larger distances that don't unduly affect the f(0) or h(0) estimates.

The weighting function we use is $\cos(\pi x_i/2w)$ where x_i is the perpendicular distance of observation i and w is the truncation distance. This leads to the statistic

$$C^2 = \frac{-16n}{\pi^3} + \frac{2}{\pi}\sum_{i=1}^{n}\left[\left(2\hat{F}(x_i) - \frac{2i-1}{n}\right)\sin\left(\frac{\pi}{2}\hat{F}(x_i)\right) + \frac{4}{\pi}\cos\left(\frac{\pi}{2}\hat{F}(x_i)\right)\right]$$

Tail probabilities for C^2 are calculated and presented in the same way as for W^2, above.

These tests are also discussed in Chapter 11 of Buckland *et al.* (2004). For options associated with goodness-of-fit tests in the CDS engine, see the Program Reference page on Model Definition Properties, Diagnostics - Detection Function Tab - CDS and MCDS.

About CDS Detection Function Formulae

Understanding the Detection Function Model Formulae

On various pages of the results details listing, Distance presents the detection function model used and the parameters estimated. For example, under **Detection Function**, **Model Fitting**, you may see:

```
Model
    Hazard Rate key, k(y) = 1 - Exp(-(y/A(1))**-A(2))
    Simple polynomial adjustments of order(s) :  4
       A( 1) bounds = ( .00000     ,  1.0000     )
       A( 2) bounds = ( 1.0000     ,  2.0000     )
```

and under **Detection Function**, **Parameter Estimates**:

```
Model
    Hazard Rate key, k(y) = 1 - Exp(-(y/A(1))**-A(2))
    Simple polynomial adjustments of order(s) :  4

              Point         Standard    Percent Coef.
  Parameter   Estimate        Error      of Variation
  ---------  -----------  -----------  --------------
    A( 1)     0.4582       0.3373
    A( 2)      1.195        1.085
    A( 3)     -1.954       0.4343
    f(0)      1.5652      0.49259          31.47
  ---------  -----------  -----------  --------------
```

Detection functions are modelled using the following general form:

$$g(y) \propto key(y)[1 + series(y_s)]$$

where $g(y)$ is the detection function, $key(y)$ is the key function, $series(y_s)$ is the series adjustment, y is distance and y_s is the scaled distance y/w where w is the truncation distance. The function is scaled so that $g(0)$ is 1. In the above example, the key function is hazard rate and the series adjustment is one polynomial term of order 4.

The formula for the hazard rate function is given in Buckland et al. (2001) as

Hazard rate: $1 - \exp\left(-(y/\sigma)^{-b}\right)$

where σ is the scale parameter and b is the shape parameter. This formula is given in the output above, and it can be seen that parameter A(1) corresponds to σ and A(2) to b.

The formula for the polynomial series adjustment is given in Buckland et al. (2001) as

Simple polynomial: $\sum_{j=2}^{m} a_j (y_s)^{2j}$

where m is the number of adjustment terms, and a_j is the parameter for the adjustment term of order $2j$. In the above output, parameter A(3) corresponds to the order 4 adjustment term (i.e., where j=2).

In the MCDS engine, y_s can also be y/σ where σ is the scale parameter of the key function - see Scaling of distances for adjustment terms in the MCDS chapter.

Calculating probability of detection

To calculate probability of detection at a given distance, y, you need to substitute the parameter estimates into the formula

$$g(y) = \frac{key(y)[1 + series(y_s)]}{key(0)[1 + series(0)]}$$

For example, taking the results given above and assuming a truncation distance $w = 10$, the probability of detection at $y = 3$ is

$$\begin{aligned} g(y) &= \frac{\left(1 - \exp\left(-(3/0.4582)^{-1.195}\right)\right)\left[1 - 1.954(3/10)^4\right]}{\left(1 - \exp\left(-(0/0.4582)^{-1.195}\right)\right)\left[1 - 1.954(0/10)^4\right]} \\ &= \frac{(1 - 0.8995)[0.9842]}{(1 - 0)[1]} \\ &= 0.0989 \end{aligned}$$

To calculate average probability of detection over the surveyed strip, an easy approach is to divide the interval (0, w) into a large number of evenly spaced intervals, evaluate $g(y)$ at each cutpoint and take the mean. (This is almost equivalent to numerical integration of $g(y)/w$ using the trapeziodal rule.) Alternatively, use a numerical integration routine (e.g., the function `integrate` in R) to integrate $g(y)$ and divide the result by w.

NEW! Tip! Because $g(y)$ is a nonlinear function of y, the accuracy of calculated average probability of detection will be much less than the 4 or 5 significant figures Distance reports for the parameter estimates under **Detection Function**, **Parameter Estimates**. For accurate estimates, it is better to use the estimated parameter values given in the stats file (to 7 significant figures). To do this, ask Distance to save the results stats file (see Saving CDS Results to File in this chapter) and then look for the statistics numbered 101, 102, etc. in module 2 (see MCDS Engine Stats File in the MCDS Engine Reference for more about the stats file format).

Calculating effective strip width

Advanced Topic

The estimated effective strip width (line transects) or effective area (point transects) in CDS analysis is given in the Results Details listing. However, there are some circumstances when you may wish to calculate it outside of Distance using the parameter estimates (one example is for MCDS analyses to calculate it at a given covariate level).

For line transects, effective strip width is given by

$$\mu = \int_0^w g(y)dy$$

where *g(y)* is the probability of detection at distance y and w is the truncation distance. Effective strip with can therefore be calculated by numerical integration of *g(y)*. Calculating *g(y)* from parameter estimates output by Distance is described in previous sections of this chapter.

For point transects, effective area is given by

$$\nu = 2\pi \int_0^w rg(r)dr$$

where *g(r)* is the probability of detection at distance r and w is the truncation distance. Effective area can therefore be calculated by numerical integration of *g(r)*.

Tip! See the previous topic for a tip on how to get the best accuracy when calculating statistics such as effective strip width (and average probability of detection).

Understanding parameter indexing

Advanced Topic

To set starting values or bounds on parameters, it is important to understand the order in which they are indexed. This can become quite complicated when there are multiple strata and/or multiple models.

- Within a stratum/model, the key function parameters come first and adjustment term parameters next. In the above example, parameters A(1) and A(2) are key function parameters and A(3) is the adjustment term parameter. Additional adjustment term parameters would be A(4), A(5), etc
- When fitting detection function by stratum (or sample), parameters are indexed sequentially between strata (or samples). For example, if there were two strata in the above example, then the hazard rate key function parameters would be A(1) and A(2) in stratum 1 and A(4) and A(5) in stratum 2, and the adjustment term parameter would be A(3) in stratum 1 and A(6) in stratum 2. You can see these parameter indexes in the **Detection Fct** part of the output.
- When fitting multiple models, the parameters are indexed separately. So, if in the Model Definition under **Detection Function**, **Models**, you specified both a hazard rate key function and a half normal key function, then the hazard rate parameters would be indexed starting with A(1) and A(2) and the half-normal parameter would be indexed starting with A(1).
- You can use the above rules to determine which parameters refer to which strata when setting starting values. When setting bounds, remember that only key function parameters are included in the ordering.

Tip! The best way to be sure which parameter is which is to run the analysis without starting values and check the **Detection Fct** part of the output.

- When there are multiple models and multiple strata, the parameters are not indexed separately when printing results in the **Density Estimates** section. Only the model selected in each stratum contributes to the indexing. For example, imagine Distance is choosing among hazard rate and half-normal key functions, with no adjustments, and that there are two strata. In stratum 1 it chooses half-normal and in stratum 2 it chooses hazard rate. Then in the **Density Estimates** output, parameter A(1) corresponds to the half-normal parameter in stratum 1 and parameters A(2) and A(3) correspond to the hazard rate parameter in stratum 2.
- This means that in some cases the parameter indexes in the **Density Estimates** part of the output can be different from those in the **Detection Fct** part. Hopefully the output in each section is self-explanatory. The important thing to remember is that it is the parameter indexes in the **Detection Fct** part that are used for setting starting values – the output in the **Density Estimates** part is for display purposes only.

CDS Analysis Browser Results

When an analysis is run, a summary of the results is given in the right-hand pane of the Analysis Browser. You can select which statistics that are displayed separately for each analysis set by using the Column Manager (click the button).

Most of the columns that are available for selection have obvious interpretations. However a few require some additional explanation or amplification:

- Many columns will appear blank in the Analysis Browser when the analysis is stratified. For example, the number of parameters, probability of detection and chi-square *p* columns will be blank if detection function is estimated by stratum.
- One exception to the above is the model selection statistics AIC, AICc, BIC, and LogL (and respective Delta AIC, Delta AICc, …). When detection function is estimated by stratum, these statistics are summed across the estimated detection functions, making it easy to compare models where detection function is estimated separately by stratum vs those where it is pooled.
- The goodness of fit Chi-square p value is for the last test performed (if more than one diagnostic test is performed), after automatic pooling has taken place - see the last "Chi-sq GOF" page of the Analysis Details Results Tab.
- Both bootstrap and analytic estimates of coefficient of variation and confidence limits for the abundance and density estimates can be displayed. The bootstrap estimates use the percentile method (cf. bootstrap in the Distance Book). Bootstrap estimates obtained from the bootstrap variance estimate, assuming a lognormal distribution for the density estimate, are available in the Analysis Details Results Tab Bootstrap Summary page.
- Bootstrap point estimates of abundance and density are also available – these are the mean of the estimates from the bootstrap replicates. They are especially useful if you have run the bootstrap with multiple key functions as they are then model-averaged point estimates – see Model Averaging in CDS Analysis for details.

Exporting CDS Results

There are many reasons to want to export the results of analyses to other programs. For example, you may want to:

- present a summary table of results from the Analysis Browser in a report
- save part of the detailed results from the Analysis Details window
- copy a detection function graph into a spreadsheet program and modify the formatting
- export the parameter estimates from Distance as a text file, as part of a simulation

This section summarizes the various ways of getting results out of Distance.

Exporting CDS Results from the Analysis Browser

To copy the current analysis set to the clipboard, click the **Copy to Clipboard** button on the main toolbar, or choose **Analyses** | **Copy to Clipboard**. The column and row separators can be set in the **General** tab of the **Preferences** dialog (**Tools** | **Preferences**).

Tip! Non-integer results are only shown in the Analysis Browser to a few (usually 2) decimal places. However, if you copy and paste them into another application (e.g., Excel) you can see them to 7 significant figures.

Exporting CDS Results from Analysis Details Results

Exporting Text

To transfer the results text, make sure a page of text is showing and then click on the **Copy to Clipboard** button on the main toolbar, or choose the **Analysis – Results** menu button **Copy Results to Clipboard**. Then paste into whatever application you choose.

To transfer just part of the result text, highlight the text, right click and choose **Copy Selected Text**.

Exporting Plots or Plot Data

To transfer a high-resolution plot, make sure the plot is showing and then click on the **Copy to Clipboard** button on the main toolbar, or choose the **Analysis – Results** menu button **Copy Plot to Clipboard**. You can only paste the plot itself into applications that are designed to take picture objects, such as word processors and spreadsheets. In your application, choose **Paste Special**, and then choose the option "Picture" from the list of formats that appears (you can also choose "Device Independent Bitmap", but you will end up with a much larger file).

Alternatively, you can paste the data that was used to generate the plot file into, say, a spreadsheet and then regenerate the file yourself. This has the advantage that you can then change the format of the plot. To do this, in your application, choose **Paste**, or choose **Paste Special** and then the "Unformatted Text" option.

Tip! If you wish to use Excel to recreate a plot from the plot data, there is a simple macro available on the Support page of the Program Distance Web Site to do this.

Tip! If you wish to use R to recreate a plot, the following instructions may help. (1) Paste the data from the clipboard to a text file. Let's say the file is "plot.txt". (2) Paste the following commands into R:

```
#this reads in the file just created
forplot<-read.table(file="plot.txt", header=T, sep="\t", dec=".")
#note, depending on your language, dec might be "," rather than "."
#this plots the detection function or pdf (if point transects)
plot(forplot$C1, forplot$C2, type="l", ylim=c(0,max(forplot$C4)),
     xlab="Distance", ylab="Detection probability")
#Define labels as you wish
#this adds in the data bars
lines(c(0,0), c(forplot$C3[1], forplot$C4[1]))
lines(forplot$C3, forplot$C4)
```

The following code can be used to recreate qq-plots:

```
#this reads in the file just created
forplot<-read.table(file="plot.txt", header=T, sep="\t", dec=".")
#note, depending on your language, dec might be "," rather than "."
#this plots the detection function or pdf (if point transects)
plot(forplot$C1, forplot$C2, type="p", ylim=c(0,max(forplot$C4)),
     pch = ".", xlab="Empirical distribution function",
     ylab="Fitted cumulative distribution function")
#Define labels as you wish
```

```
#this adds in the (0,0) (1,1) line
lines(c(0,1), c(0,1))
```

and the labels should be changed.

Saving CDS results to file

In the Model Definition Properties dialog, there are options to save files containing summaries of the results of an analysis. Four files can be saved, all of which are standard ASCII text files:

- Results Details File - see **Misc**. Tab - CDS and MCDS.
- Results Stats File - see **Misc**. Tab - CDS and MCDS.
- Bootstrap Stats File - see **Variance** Tab - CDS and MCDS.
- Plot File - see **Detection Function Tab** - CDS and MCDS.

These files may be useful in providing an interface between Distance and other applications - for example you could write a spreadsheet macro to paste the results stats file and extract information into spreadsheet cells. In addition, the Bootstrap file is often useful for making diagnoses of problems encountered while doing bootstrap resampling. The formats of these four files are given in the MCDS Engine Command Language Appendix section Output from the MCDS Engine.

Note that the results details file is the same as the text displayed in the Results tab of the Analysis Details window for an analysis that has been run. You can easily obtain this text by choosing the menu item **Analysis – Results | Copy Results to Clipboard** or pressing the the **Copy to Clipboard** button on the main toolbar. Similarly, you can obtain a copy of the plot data by displaying the plot in Distance and pressing the **Copy to Clipboard** button. For more on this, see Exporting CDS Results from Analysis Details Results.

Miscellaneous CDS Analysis Topics

Interval (Binned/Grouped) Data

Summary

If your data were collected in intervals (bins), then enter them as exact distances and convert them to intervals in the **Intervals** tab of the **Data Filter**.

Details

In Distance, each record in the Observation data layer corresponds to one observation, and this observation must be entered as an exact radial or perpendicular distance. In reality, however, some surveys collect distance data in distance intervals (bins, also called grouped data).

To enter these into Distance, enter each observation at the mid-point of the interval. For example if your intervals span 0-10m, 10-20m and 20-50m then an observation in the first bin would be entered as 5m, in the second bin would be 15m and in the third bin would be 35m.

If you observed, say 50 objects in the first bin of a transect you won't want to have to create 50 records by hand and type the distance for each one. Luckily there is a shortcut - simply create a record for the first object and enter

the distance (in this case 5m). Then double-click on the ID field to bring up the multi-record add dialog. Select 49 and press append - this will automatically add the other 50 records. See the Program Reference page Editing, Adding and Deleting Records for more details.

When you have entered your data, the first thing you should do is tell Distance to turn the data into intervals for analysis. Create a new analysis in the Analysis Browser, and in the Analysis Details window, click on **Properties…** for the default Data Filter. In the Intervals tab of the Data Filter click on "Transform distance data into intervals for analysis" and enter your intervals.

Look at the Data Filter Truncation tab help page to find out about choosing the level of truncation for your distance data.

Missing Data in CDS Analysis

Missing Distances

The conventional distance sampling engine is not designed to deal explicitly with missing distance data. With good field methods, it should be very rare that you detected an object but did not record a distance for it. However, if this does occur you have two options:

- Discard the observations with missing distances and analyze the data as usual. So long as the observations were not at zero distance, this should cause no bias, as you are effectively just making the detection function steeper by discarding some observations that were made away from the line or point. A disadvantage of this approach is that you are discarding data that could help to estimate encounter rate, so the overall variance may be higher than the second approach.
- Use a data filter to select only the observations with recorded distances. Fit a detection function to these data, and record the estimated probability of detection and SE. Enter these as multipliers and the estimate density/abundance using the whole dataset, as described under Multipliers in CDS Analysis. Note an important assumption here is that the missing distances are missing at random – for example it will not work if you are less likely to record the distance for objects farther from the line or point. For this reason, the first approach is probably safer. Note also that this approach won't work if you use stratification – in that case you'll need more than one multiplier (one for each stratum), and will have to calculate the global density estimate by hand.

If you run an analysis with data that includes missing distances, the CDS engine will issue a warning and exclude the observations with missing distances from the analysis.

Missing Cluster Sizes

See Missing Cluster Size Data in CDS Analysis.

Missing Survey Effort

Nothing can be done about this – to estimate density you need to know how much survey effort you expended!

Missing Study Area Size

See Unknown Study Area Size.

Clusters of Objects

In many studies, the objects of interest (usually animals) occur in clusters (schools, flocks, etc.). In this case, each observation represents a cluster, and in addition to the distance from the transect to the cluster the observer also records the cluster size.

Distance sampling theory is readily extended to include clustered populations, as outlined in Buckland et al. 2001. Distance allows you to specify that objects are in clusters during the New Project Setup Wizard. It then automatically creates a field for cluster size in the Observation data layer, and specifies that objects are clusters in the default Survey object.

The key decision in the analysis of clustered data is how to estimate the expected cluster size at zero distance. Distance offers a number of options for this, as explained in Section 3.5 of Buckland et al. 2001. In the Data Filter Properties, under Truncation, there is an option to right-truncate the data for cluster size estimation independently of the truncation for estimating the detection function. In addition, in the Model Definition Properties, there is a Cluster Size tab which gives a number of options for estimating expected cluster size (see the Program Reference page on the Cluster Size Tab - CDS and MCDS).

Cluster size is usually an integer value, but Distance allows you to enter non-integer cluster sizes. This may be useful, for example, if cluster size is estimated independently by several observers – you could then improve accuracy by using the mean of these estimates.

Missing Cluster Size Data in CDS Analysis

In some cases, you may not know the cluster size of an observation. For example, some shipboard line transect surveys of cetaceans operate in two survey modes: "passing mode" where observers guess the cluster size for each sighting and "closing mode" where they break of the survey and go to an observation to get a confirmed school size. One protocol would be to go into closing mode for every 5th sighting, say. You may want to do an analysis on only the confirmed school sizes, treating the other observations as having missing cluster sizes.

To tell Distance that a cluster size value is missing, enter a value of -1 for that cluster size. Observations with a cluster size of -1 are included in estimating the detection function and encounter rate, but are excluded from estimation of expected cluster size. If there are any -1 cluster size values in the data, Distance issues a warning.

In previous versions of Distance, missing cluster sizes were coded as 0. Zero cluster sizes are now no longer treated as missing, but are analyzed (see next topic).

Zero Cluster Sizes in CDS Analysis

There are some situations where you may observe clusters of size zero. For example, imagine that you are surveying for a species of parasitic plant that only occurs in a certain tree. You survey by walking a line transect looking for trees, and if you find one you approach it and count the parasitic plants in the tree. One way to enter these data would be by recording the distance to each tree, and the number of parasitic plants in each tree as the cluster size. You may find no plants in a tree, in which case you record a cluster size of zero. In this case, you would not want to use a size bias method to get the expected cluster size, but

instead would use the mean cluster size (see Cluster Size Tab - CDS and MCDS in the Program Reference).

An alternative analysis method would be to only include the trees where one more parasitic plant was seen in the data – cluster size will then always be >1. A disadvantage of this approach is that there will be less data available to fit the detection function.

Another example of zero cluster sizes are in multi-species analyses where species are encountered as mixed groups. You may have one field giving cluster size for one species and a second field giving cluster size for the second species. To do the analysis you would then a separate Survey object for each species and use one survey object for each analysis (see Analysis with Multiple Surveys in Chapter 7 for more on the use of multiple survey objects).

Stratification and Post-stratification

Advanced Topic

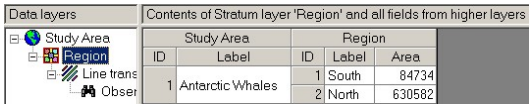
Stratification is a useful way of handling heterogeneity in the survey data, of improving precision and reducing bias (see Distance Book, Section 3.8). Stratification might be carried out by geographic region, environmental conditions, cluster size, time, animal behaviour, detection cue, observer, or many other factors.

The CDS engine can analyze data with one level of stratification per analysis. There are two ways to deal with the stratification in Distance: (1) using the stratum data layer and the **Model Definition stratify** option, and (2) using extra data fields and the **Model Definition post-stratification** option. These two methods are discussed in detail below.

Implementing Stratification via the Stratum Data Layer

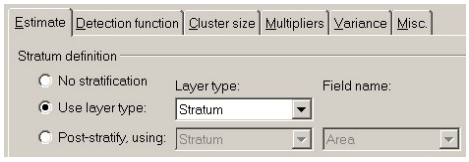
This is the recommended approach when the strata are geographic. For example, imagine a line transect study in which the study area has been stratified into two separate regions: a small area of ideal habitat and a larger area of marginal habitat. Animal density is expected to be higher in the ideal habitat and so it is given a higher density of transects than the marginal habitat (see Buckland et al. 2001, Chapter 7).

In Distance, the two regions are entered as two records in the Stratum data layer (called "Region" in this example):



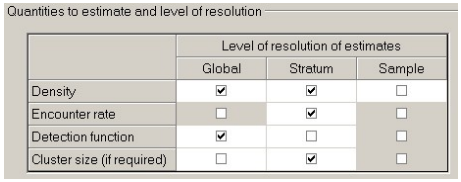
Data explorer, showing two records in the stratum layer "Region"

To analyze data that has been entered this way, you should click on the option **Use layer type: Stratum** in the **Model Definition Properties** dialog **Estimate** tab:



In the lower part of the **Estimate** tab, you can then select the level of estimation for density, encounter rate, detection function and cluster size (if the observations are clusters of individuals). In the following picture, density is estimated separately for each stratum as well as overall (globally), encounter rate and cluster size are estimated by stratum, but detection function is estimated

globally (i.e., pooled across strata). You can see this from the location of the tick marks in the boxes.



Tip! Let's assume you have no a prior reason to believe that detection function is the same across all strata. Let's also assume that you have plenty of detections in each stratum. Then you can try an analysis where you allow the detection function to vary by stratum (by ticking the "Stratum" box for detection function), and try another where the detection function is fit to data pooled across strata (i.e., estimated globally). Look at the goodness-of-fit statistics in the **Results** tab of the **Analysis Details** windows, and compare the detection function histograms. Assuming both stratum and global models produce reasonable fits, the one with the lowest AIC is to be preferred. You need to use the same **Data Filter** for both models, otherwise the AICs may not be comparable. See the Stratify example sample project for an example. This topic is discussed in Buckland et al. 2001, section 3.8.

Assuming you wish to estimate density by stratum and globally, you must tell Distance how to combine the stratum estimates to produce a global estimate. If your strata are geographic, the following options should be used:

The other options for the global density estimate are discussed in the following sections.

Implementing Stratification via the Post-stratification Option

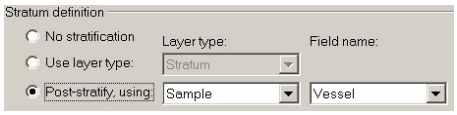
This is the recommended approach to non-geographic stratification. For example, imagine a shipboard line-transect survey conducted using two different vessels. The vessels are assigned to transects at random, but it is known that there are large differences in the effective strip width achieved by the two vessels: one ("Beagle") has a high crows-nest from which observers can see long distances, while the other ("Bounty") has no such raised platform. Distance is quite robust to such heterogeneity in detection function (see Buckland et al. 2001 and Buckland et al. in prep), but nevertheless, gains in precision may be made by modeling the heterogeneity. Therefore, the biologist wishes to allow for differences in detection function by vessel.

In Distance, the vessel information is entered as an additional field called "Vessel" in the Sample data layer (called "Line transect" in this example):

Study area		Region			Line transect			
ID	Label	ID	Label	Area	ID	Label	Line length	Vessel
1	Study area	1	Study area	1677875	1	1	251	Beagle
					2	2	235	Bounty
					3	3	432	Beagle
					4	4	354	Beagle
					5	5	365	Bounty
					6	6	234	Bounty
					7	7	341	Bounty
					8	8	254	Beagle
					9	9	324	Beagle
					10	10	266	Beagle
					11	11	315	Bounty
					12	12	402	Bounty

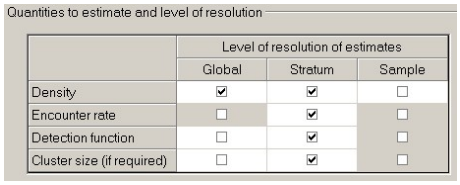
Data sheet part of the data explorer, showing the Vessel field in the sample layer "Line transect"

(For more information about how to create additional fields in Distance, see the Program Reference page about the Data Explorer. Additional fields such as this can also be imported into Distance as with the other survey data - see Chapter 5 of the Users Guide on Data Import. To analyze data where the stratum is entered as an additional field, click on the **Post-stratify, using** option in the **Model Definition Properties** dialog **Estimate** tab, and select the appropriate data layer and data field:



Example from Estimate tab, showing post-stratification by the Vessel field in the sample layer

In this example, we want to estimate detection function by stratum (i.e., by Vessel), so we fill in the **Levels of Estimation** options as follows:



Finally, we estimate the global abundance estimate as the mean of the stratum estimates, but in this case we weight by survey effort.

When you choose to weight by survey effort, the tick box **Strata are replicates** is enabled. In this case we do not want to tick that option. This, and the other options, are discussed more in the next section, and in the Program Reference page on the Model Definition **Estimate Tab - CDS and MCDS**.

Note! The CDS engine only supports one level of stratification. This means that you cannot use both geographic strata and Post-stratification at the same time. If your survey design includes geographic stratification and you create a Model Definition that defines a post-stratum field then the geographic stratification is ignored.

If you wish to estimate overall density for multi-strata surveys, you can use the **Data Selection** option in the **Data Filter** to set up a separate filter for each level of your highest stratum. Then, do a separate analysis for each level and combine the resulting density estimates by hand, remembering to include the appropriate weightings. Variances can be calculated using the delta method (Buckland et al. 2001). We hope to include multi-level stratification in a future release of Distance.

Situations where the Post-stratification option is useful

This section gives various post-stratification scenarios and outlines the way you would set the data up and the options you would choose under **Quantities to Estimate and Level of Resolution** on the **Estimate** tab of the Model Definition. More on these options is also given in the **Estimate** tab page of the Program Reference.

1. Where there are different types of survey effort.

In these situations you create an extra field that indexes the type of effort. You post-stratify on this field and estimate density as the mean of the post-stratum estimates, weighted by survey effort:

One example is the scenario described in the previous section, where there is observer heterogeneity. In this case, you would add an extra field in the sample data layer for observer (/vessel/survey party/etc.), and then post-stratify by this field. The global density estimate is given as the mean of the stratum estimates, weighted by survey effort.

Another example is where the study area is surveyed in multiple time periods, using a different set of samples (transects) in each time period. Even without wanting to post-stratify, it would be a good idea to add an extra field that indicates the time period of each transect in the sample data layer, as this would enable you to use the Data Selection feature of the Data Filter to pick out only certain time periods for analysis. Post-stratification becomes useful if you want a combined estimate of the average density over all periods. This is done by post-stratifying on the time period field, and asking for a combined estimate of density that is the mean of the post-stratum estimates, weighted by survey effort. If observers, methods and conditions were the same at all time points it would be reasonable to investigate the possibility of pooling the detection function over the time periods.

A third example is similar to the previous one, in that the study area is surveyed at multiple time points, but in this case the same set of samples (transects) were used at each time point. You could set up the project in the same way as for the previous case, but this would necessitate entering each transect into the sample data layer once for each time period. Instead, you may consider adding the time period field to the Observation data layer - this way each transect only has to appear in the sample data layer once, while you indicate the time period that each object was observed. To estimate mean density over the whole study, you post-stratify on the time period column in the observation data layer and estimate overall density as the mean in the post-strata. One small problem occurs in this scenario when you come to estimate variance - each stratum in each year will be treated as independent when in fact they are not. We hope to address this in a future release of Distance.

Strata as replicates when there are different types of survey effort

In all of the above cases, we chose to weight by survey effort. When you weight by survey effort, there is an option to treat the strata as replicates. This affects how the variance of the global density estimate is calculated. Ticking the **Strata are replicates** option means that you consider the strata you surveyed to be a random sample from some larger population of possible strata that could have been surveyed.

For example, consider the case when we survey at multiple time points – say multiple days during a year. We may consider the days we surveyed to be a sample from all those in the year, and we want to make inferences about the average density of animals during the year. In this case we tick the **Strata as replicates** option. Our estimate is then the effort-weighted mean density, and our variance is calculated from the variation in density between days – i.e., treating strata as replicate samples.

On the other hand, let's imagine that our multiple time points are actually years, and we only surveyed in two years (we pooled the data within year). We could consider the two years of data to be a sample from some larger set of possible years – but this does not seem very useful and in any case a sample of two is not very large for making inferences about average density over this larger set of years. Instead, we will make inferences only about the average density over the two years we sampled. We do not tick the **Strata as replicates** option. Our

estimate is still the effort-weighted mean density, but now the variance is calculated from a weighted average of the stratum variances.

The difference between the two scenarios is known in the statistical literature as treating the strata as random effects (the first scenario) or fixed effects (the second scenario). The right one for your study depends on the inferences you are making – is it to the average density over a larger set of strata from which you have a random sample (random effect), or is it just to the average density over the strata in which you sampled (fixed effect).

The variance calculation is given in more detail in the Program Reference page on the Model Definition **Estimate Tab - CDS and MCDS**.

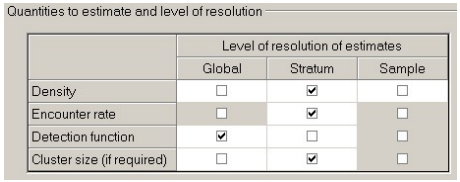
2. Where there are different types of object (animal) in the population.

If the population can be divided into different "sub-populations", each with different encounter rates or detection functions, then it may be possible to increase precision through post-stratification. This is done by creating an extra field in the Observation data layer that indexes the sub-population type. You then post-stratify on this field, with the global density estimate as the sum of the post-stratum estimates:

One example of this would be where male and female animals have very different detectabilities. For each animal, its sex would be entered as an additional field (of type "Other") in the Observation data layer. In the Model definition, you would choose Post-stratification by the sex field.

3. Where there is not enough data to estimate a detection function for some subsets of the study.

For example, in a multi-species study it is often not possible to estimate *f(0)* reliably for the rarer species. In this case, it may be acceptable to estimate the detection function by pooling over similar species. To do this, you would add a column to the Observation data layer for species name or ID. Then, define a Data Filter that uses the Data Selection to include only the species for which you wish to pool the detection function. In the Model Definition, post-stratify by species and choose the following Levels of estimation:



You are not interested in obtaining a global estimate of density, as density pooled over species is not meaningful.

To test your assumption about the detection function being similar among species, it may be possible to do an analysis with detection function estimated by species (by "stratum") and to compare the AICs, although in this scenario there are often too few observations for the rare species to make reliable estimates of the detection function.

This situation can also be addressed using Multipliers - see the topic on Multipliers in CDS Analysis.

Variance Estimation in CDS

The methods of variance estimation in CDS analysis are detailed in Buckland *et al.* (2001), section 3.6. The default method is an analytic variance estimate, but it is also possible to select a nonparametric bootstrap. A summary of the

methods used is given here; this can be safely skipped on first reading through this manual.

Variance options are chosen in the **Model Definition**, **Variance** tab – for more on these options, see Variance Tab - CDS and MCDS in the Program Reference appendix.

Analytic variance estimation

Density can be estimated globally, by stratum and/or by sample. At the lowest of the levels requested (e.g., at the stratum level of both globally and by sample are selected), variance estimates for encounter rate, detection probability and expected cluster size are combined using the delta method (formula 3.68 of Buckland *et al.* 2001) to give the variance of the density estimate at that level. Lognormal confidence intervals are calculated using formulae 3.71-3.74 except that t-based limits are calculated using degrees of freedom calculated using the Satterthwaite method given in formula 3.75.

If there are any multipliers (see Multipliers in CDS Analysis) with non-zero variance these are included in formula 3.68 as extra terms. If they have a non-zero degrees of freedom they are also included in the degrees of freedom calculation of equation 3.75. If degrees of freedom for the multiplier is not specified, it is assumed zero, and the multiplier is omitted from both the top and bottom line of equation 3.75.

By default, encounter rate variance is calculated using the empirical between-sample variation in encounter rate, as detailed in section 3.6.2 (formulae 3.77-3.82). Alternatively, the user may specify that encounter rate variance follows a Poisson or overdispersed Poisson distribution (see Variance Tab - CDS and MCDS). In this case, the encounter rate variance is assigned zero degrees of freedom and the encounter rate term is omitted from both the top and bottom line of equation 3.75.

When there is only one line, the default option is to set encounter rate variance to zero – see Analysis of Data from a Single Transect in CDS.

When density is requested to be estimated at multiple levels (e.g., by stratum and globally) then the higher density estimates are calculated as a weighted average of the lower estimates (see Estimate Tab - CDS and MCDS in the Program Reference for how to specify what weighting to use). In this case, the formulae in section 3.7.1 are used to calculate variance of the estimate at the higher levels – details of exactly which formula is used in which circumstance are given in the Program Reference under Estimate Tab - CDS and MCDS.

Bootstrap variance estimation

The bootstrap is a robust procedure for estimating variance and confidence limits (among other things). Details of the bootstrap implemented in the CDS engine are given in Buckland et al. (2001), section 3.6.4, and a description of how to tell Distance to produce bootstrap estimates is in the Program Reference appendix under Variance Tab - CDS and MCDS. Note that the bootstrap can also be used for producing estimates averaged across a number of candidate detection function models – for more on this see Model Averaging in CDS Analysis in this chapter.

Multipliers in CDS Analysis

In Distance sampling, there are often situations where the standard methods produce a density estimate that is only proportional to the true density. For example, when detection probability on the trackline (g0) is less than one, the

true density is the Distance density estimate divided by g0. Another example comes from cue counts, where the density of animals is the density of cues divided by the cue rate.

In Distance, such factors are dealt with using Multipliers. You enter the multiplier value in the global data layer and in the **Multipliers** tab page of the Model Definition Properties, you tell Distance which multipliers you want to use in your analyses and whether they divide or multiply the density estimate. Distance then scales you density estimate appropriately.

Some multipliers are known with certainty. One such multiplier is the "sampling fraction" – the proportion of each line or point surveyed. Normally this is 1, but in some cases you may only survey one side of a transect line, so the sampling fraction is 0.5. Another example is a cue count survey, where the sampling fraction is the proportion of a full circle that is covered by the observation sector. A third example is when all points or lines are visited multiple times – then the sampling fraction is the number of visits. In these cases, you tell Distance to create a field for the multiplier, enter the appropriate value, and tell Distance to multiply the Density estimate by the value in this field.

If the sampling fraction is not the same for all lines or points, you account for this by adjusting the survey effort at the data entry stage. For example, if all your transects were 10km long, but you visited some 3 times, then you set the survey effort for these to 30km, and leave the others at 10km. In this situation you don't need a sampling fraction multiplier.

Other multipliers are based on estimates from other experiments. For example, in a cetacean survey, g(0) may be less than one because some animals are below the surface and so not available for detection. You may have estimated g(0) based on a separate experiment where you follow a sample of animals and record the proportion of time they are on the surface. In these cases, your estimate of the multiplier will have uncertainty associated with it. If you want this uncertainty to be reflected in the variance of the final density estimate, you do this in Distance by having additional fields for the multiplier standard error (SE) and degrees of freedom (df). Note that you can have fields for both SE and df, or just the SE. In this latter case, Distance assumes the DF for the multiplier is infinity. (Another way to specify infinite DF is to have a field for DF containing the value 0.0) Degrees of freedom of the multiplier affect the DF of the density estimate as the density estimate DF is calculated using the Satterthwaite approximation (formula 3.75 of Buckland et al. 2001). This in turn affects the log-normal confidence limits on the density estimate.

In the Setup Project Wizard you are given the chance to create fields in the Global data layer for a number of common multipliers. You can also add more fields manually after the project is created using the Append Field button in the Data Explorer (see Data Explorer in the Program Reference).

In this version of Distance you can only define Multipliers in the global data layer. This means that you cannot account for factors that vary by stratum or transect. In future versions of Distance we hope to offer the ability to define multipliers at any level of the data hierarchy.

If you use the Setup Project Wizard to define your multiplier fields, then they will appear automatically in the Multiplier tab in Model Definition Properties for the default CDS analysis. For these fields, Distance also knows whether they should multiply or divide the density estimate. The default value for a multiplier created by the wizard is 1.0, with SE 0 and DF 0 (i.e., infinity) –

in other words a multiplier that doesn't affect the density estimate at all! It's up to you to enter appropriate values into the multiplier fields.

It is a good idea to follow Distance's convention for naming your multiplier, SE and DF fields. If the multiplier field name is, for example, Nest Prod Rate (for Nest Production Rate in indirect surveys of nests), then the corresponding SE field should be Nest Prod Rate SE and DF field should be Nest Prod Rate DF. This makes it much easier to recognize which SE and DF goes with which multiplier.

If you are estimating variance using the bootstrap, be aware that the variance due to any multipliers is *not* included in the bootstrap variance. To do this, we would have to resample the multiplier value in each bootstrap iteration, according to some distribution – a possible feature to add to a future version of Distance…

In fact, the CDS analysis engine only recognizes the multiply operator - i.e., all multipliers multiply the density estimate. So, the Distance Interface uses a trick to pass multipliers to the Analysis Engine when the operator is divide. In this case, to pass the multiplier value, it takes 1 / (input value). To pass the multiplier SE, it takes SE / (input value^2). You can see this happening if you look at the first page of the Analysis Details Results tab for an analysis that has been run with multipliers, where the operator is a divider. About half-way down the first page is a small table listing the multiplier values and SEs used. If, for example, you used a multiplier with a value of 0.9 and an SE of 0.5, you'll see that the actual value passed in was 1.1 (i.e., 1 / 0.9) and the SE was 0.6128 (i.e., 0.5 / (0.9^2)).

Additional Uses of Multipliers

This section lists an example of another use for multipliers. If you think of any more, let us know and we'll include them in future versions of the help file.

In a multi-species study it is often not possible to estimate detection probability reliably for the rarer species. In this case, you may want to assume that the probability of detection for a rare species is the same as that for a similar, more common one. You can do this using multipliers:

8. Define a Data Filter that uses the Data Selection tab to include only the more common species. Perform a normal distance analysis and note down the estimated P (detection probability) and P SE (standard error) from the "Density Estimates/Global" page of the Analysis Details Results.
9. In the Data Explorer, create a new multiplier, and multiplier SE and enter the estimated P and P SE in the cells. (You can do this using copy and paste - try highlighting the value in the results page and clicking the right mouse button).
10. Now, define a Data Filter that uses the Data Selection tab to include only the rare species of interest. Define a Model Definition that uses a Uniform key function (under Detection Function Models) and 0 adjustment terms (under Detection Function Adjustment Terms). Include the new multiplier and multiplier SE in this Model Definition under the Multipliers tab.
11. When you run the analysis, Distance will fit a uniform distribution to the detection function with no adjustments and so will estimate

probability of detection as 1.0. The multiplier you created will allows you to specify the probability of detection yourself. Variance of the density estimate is calculated correctly automatically.

Another approach to this scenario is to estimate a pooled detection function for both the common and rare species. This is outlined in the previous page on post-stratification.

Another example of the use of multipliers is given in the Getting Started chapter, Example 2: More Complex Data Import.

A third example would be when some distances are missing from the dataset, and you are confident that these are missing at random (i.e., it isn't just the farthest away distances that are missing). Then you could fit a detection function to the data for which you have distances, enter the estimated detection probability (and associated SE and df) as a multiplier and estimate density using the whole dataset. For more on this, and a perhaps better approach, see Missing Data in CDS Analysis.

Model Averaging in CDS Analysis

When using AIC to select among alternative candidate models of the detection function, it is not unusual to find that more than one model have similar AIC scores (perhaps differing by AICs of 2 or fewer). When this happens, more reliable inferences can be obtained by basing the final results on an AIC-weighted average of these plausible alternative models (Buckland et al. 1997; Burnham and Anderson 2002). To do this:

- Create a new Model Definition.
- In the **Detection Function**, **Models** tab, click on the **+** button to create one line for each candidate model. For each one, set the appropriate key function and adjustment terms. Select the option to **Select among multiple models** using AIC
- In the **Variance** tab, tick on **Select non-parametric bootstrap**, and set appropriate options for the level of bootstrapping and number of bootstraps.

When you run an analysis using this model definition, each bootstrap replicate will use AIC to choose among the candidate models. The bootstrap point and interval estimates you get will then be an average over all the replicates, and so will include uncertainty as to which model is best.

You can only use this approach to include different candidate key function + adjustment term combinations. There is currently no way within Distance to do model averaging over, say, global and by-stratum analyses, or CDS and MCDS models, or different covariates within an MCDS model. This would have to be done by writing an external bootstrap routine and running the analysis engine as a stand-alone program (see Running CDS Analyses From Outside Distance).

For more on the bootstrap options, see Variance Tab - CDS and MCDS in the Program Reference.

Sample Definition in CDS Analysis

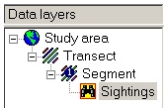
Advanced Topic

Sample definition allows you to specify the data layer to be used in the estimation of the encounter rate variance.

For example, imagine you have line transect data from a survey where 3 lines are divided into 4 short segments. The spacing between segments is approximately the same as the spacing between transects, so a map of the design is as follows:

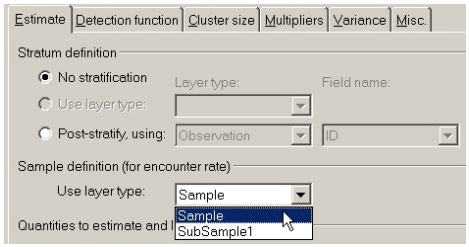
- - - -
- - - -
- - - -

The data are stored in distance in a Global layer called “Study area”, a Sample layer, “Transect”, a Sub-sample1 layer, “Segment”, and an Observation layer, “Sightings”:



Data layers view from the example project

When you come to analyze these data, you must choose whether to treat the 3 transects or the 12 segments as the samples, for estimating variance in encounter rate. You make this selection in the **Sample definition** section of the **Estimate** tab of the **Model Definition Properties**:



Selecting the layer type to use as sample

In this case, because the distance between segments is the same as the distance between transects, it is valid to treat each segment as a separate sample. So you would choose the layer type “SubSample1” as the sample definition.

If the between-segment spacing was much less than the between-transect spacing, then you would choose the layer type “Sample” as the sample definition, and Distance would pool the data from the segments on each transect for estimating encounter rate variance.

Unknown Study Area Size

If you don't know the size of the study area you sampled, you can obtain density estimates from Distance, but not abundance estimates.

When entering your data, leave the area of each stratum at its default value of 0. When you come to analyze the data, Distance will automatically detect that the area is 0 and will not provide an abundance estimate.

Restricting Inference to Density or Abundance in the Covered Region in CDS Analysis

Normally in distance sampling, we lay down a series of line or point samplers within some large study area and use the observations from these samplers to

make inferences about the density or abundance of animals in the study area. However, in some (rare) cases, we may just want to make inferences about the density or abundance of animals in the region covered by the samplers. One example where we actually survey the whole study area – for example in a simulation experiment, where we lay out objects such as golf tees or wooden stakes in a strip and then have observers do a line transect experiment within that strip. Another example is when we do not lay out the transect lines using a survey design with an element of randomization in it, but lay the lines purposively. In this case, we cannot guarantee that our survey lines will be representative of the study area and so one approach is to restrict inferences only to density in the region actually covered by the lines. (Another is to use a model-based approach such as that of Hedley *et al.* 2004)

Restricting inferences in this way does not affect the density estimate at all. Abundance is simply density multiplied by the area covered (the sum of the areas of the samplers). The big difference comes in the variance. Normally in distance sampling, there are three components that make up the variance of the density or abundance estimate: variance from estimating the detection probability, variance from estimating population mean cluster size and variance from spatial variability in encounter rate between samplers. The third component is often the largest, and typically it is estimated from the empirical variation in encounter rate between samplers. When we only wish to make inferences about density or abundance in the covered region, the only relevant sources of uncertainty are the first two – the third is no longer relevant.

If we wish to restrict inferences in this way, how can we ensure that Distance sets the encounter rate variance to zero? The trick is as follows. For the CDS (and MCDS) engine, in the **Model Definition**, **Variance** tab, under **Analytic variance estimate**, choose the option **Assume distribution is Poisson, with overdispersion factor** and set the factor to 0. Once you've run the analysis, you can check in the results that the encounter rate variance is zero. If you're only interested in density, this is all you have to do. If you also want the correct abundance estimate, you need to set study area (either in the global layer or stratum layer, if you have strata) to be the sum of the area of the samplers (either globally or by stratum) – i.e., the total length times twice the truncation width for line transects or number of points times truncation radius for point transects.

Analysis of Data from a Single Transect in CDS

In general, it is bad practice to try to estimate density in some large study area (or stratum) using just a single transect (see, e.g., Buckland *et al.* 2001 section 7.2). When data come from a single transect, it is not possible to use the default method to estimate encounter rate variance – that is using the empirical between-line variation in encounter rate. Instead, the CDS (and MCDS) engine issues a warning, and assumes that encounter rate variance is zero. The resulting variance is appropriate if you only wish to make inferences about the density or abundance of animals in the area actually sampled – possibly a wise choice when there is only one line. For more about this see the earlier section on Restricting Inference to Density or Abundance in the Covered Region in CDS Analysis.

An alternative, if you wish to try to make inferences about the whole area from one line, is to assume the distribution of observations is a Poisson or overdispersed Poisson (an overdispersion factor of 3 has been suggested in a related context by Buckland *et al.* 2001: section 7.2.2). You select these options in the Variance tab of the Model Definition – see Variance Tab - CDS and MCDS in the Program Reference appendix.

Running CDS Analyses From Outside Distance

The CDS engine is implemented as a stand-alone FORTRAN program, MCDS.exe. This program is called behind the scenes by Distance when you press the **Run** button on the Analysis Details Inputs tab.

Some users may wish to run the engine from outside the Distance interface – either from the Windows command line or from another program. For example, you may want to automate the running of analyses for simulations, or you may want to perform a more complicated bootstrap than Distance allows.

Full documentation for running MCDS.exe is provided in the Appendix - MCDS Engine Reference.

Chapter 9 - Multiple Covariates Distance Sampling Analysis

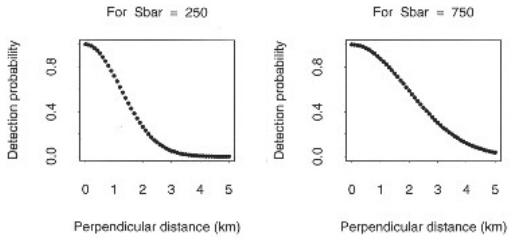
Introduction to MCDS Analysis

This chapter gives an outline of multiple covariates distance sampling (MCDS) in Distance, together with guidelines for approaching MCDS analyses. The methods are described in more detail in Marques (2001), Marques and Buckland (2003) and Marques and Buckland (2004) and Marques et al. (2007 – see Bibliography). These last two texts are recommended reading to anyone using these methods. The dataset analyzed in Marques et al. (2007) is included with Distance as a sample project (see the Sample Projects page of Chapter 3, Amakihi project). Note that MCDS is an advanced topic that should only be considered once you are familiar with conventional distance sampling.

The standard method for performing MCDS analysis in Distance is by using the MCDS Engine, as introduced in the next topic Introducing the MCDS Engine. It is also possible to perform an MCDS analysis using the MRDS engine – see Single Observer Configuration in the MRDS Engine in Chapter 10 of the Users Guide.

Introducing the MCDS Engine

The multiple covariates distance sampling (MCDS) engine extends the key function + adjustment term approach of the CDS engine, allowing additional covariates into the scale parameter of the key function (via a log link function). This means that the covariates are assumed to influence the *scale* of the detection function, but not its *shape* (see figure, below). In other words, we are assuming the covariates affect the rate at which detectability decreases with distance, but not the overall shape of the detection curve.



Example estimated detection functions, where cluster size (Sbar) is the covariate. The basic shape of the function is the same (half-normal), but the effective strip width is wider at cluster size 750.

Note! Of course, it is possible for covariates to affect both the shape and scale of the detection function. Such models could be fit in Distance (for factor covariates) using stratification. There is, however, some evidence that at least some covariates may only affect the scale: Otto and Pollock (1990) examined the effect of cluster size and distance on the detection function of graduate students searching for beer cans. They found that a model where cluster size influenced only the scale of the detection function fit the data best.

Note! When adjustment terms are included in the model, it is possible for covariates to influence both the scale and shape of the fitted function – see the section Scaling of Distances for Adjustment Terms further on in this chapter for more information about this.

Aside! The MCDS engine is implemented as a FORTRAN program which is run from within the Distance interface. You can also run the engine as a stand-alone program – for details see the Appendix - MCDS Engine Reference.

Setting up a Project for MCDS Analysis

Before you contemplate an MCDS analysis, you should be familiar with the setup and analysis of CDS data (see previous chapters).

Covariates are simply entered or imported as extra fields of data, in the appropriate data layer. For example, covariates such as habitat or observer might be associated with each point transect, and so be entered as extra fields in the sample layer, with one record for each point. Alternatively, covariates such as the sex or species of detected animals might be associated with each observation, and so be entered as extra fields in the observation layer.

Then, for an example of importing data containing an extra field for species, see Getting Started Example 2: More Complex Data Import in Chapter 3.

Defining MCDS Models

This section describes how MCDS models are defined. More details are given in Marques and Buckland (2001, 2004) and Marques et al. (2007).

Like the CDS engine, the MCDS engine uses a key function + series expansion formulation to model the detection function. The difference is the incorporation of covariates in addition to distance into the key function. So,

$$g(x,\mathbf{z}) = \text{key}(x,\mathbf{z})[1 + \text{series}(x)]$$

where $g(x,\mathbf{z})$ is the probability of detecting an object at distance x and covariates $\mathbf{z}$, $\text{key}(x,\mathbf{z})$ is the key function, and $\text{series}(x)$ is the series expansion.

The covariates are assumed to affect the scale parameter of the key function, σ. The scale parameter controls the "width" of the detection function. Of the four key functions available in the CDS engine, the half-normal and hazard-rate are both available in the MCDS engine; the other two either do not have a scale parameter (uniform), or provide an implausible shape close to 0 distance (exponential).

Half-normal key function, $\exp\left\{-x^2/2\sigma(\mathbf{z})^2\right\}$

Hazard-rate key function, $1-\exp\left\{-\left[x/\sigma(\mathbf{z})\right]^{-b}\right\}$

The scale parameter is modeled as an exponential function of the covariates:

$$\sigma(\mathbf{z}) = \exp\left(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots \beta_q z_q\right)$$

where q is the number of covariate parameters. The term inside the brackets is akin to a linear model – the β's are parameters to be estimated, with β_0 corresponding to the intercept. The exponential term prevents the scale parameter from being negative.

Output in distance is actually given with first parameter, β_0, outside the exponential – i.e.,: $\sigma(\mathbf{z}) = \beta_0 \exp\left(\beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_q z_q\right)$ This is because in this formulation, β_0 parameter estimates can be directly compared with estimates of the scale parameter from the CDS engine.

Factor and Non-factor Covariates in MCDS

When setting up MCDS models in Distance, you need to distinguish between **factor covariates**, and **non-factor covariates**. This is because the type of covariate affects how the model is parameterized. Parameterization in the MCDS engine is briefly outlined below, but will be covered in more detail in any standard book on linear models.

Factor covariates classify the data into more than one distinct category or level. Examples include habitat, sex, observer, etc. For factor covariates, the actual covariate values are not important, as a separate parameter is defined for each factor level. The last factor level is incorporated into the intercept. For example, imagine a habitat covariate with 3 levels: "grassland", "scrub", "wood" (factor covariates are sorted into alphabetical order when run in Distance). The model for the scale parameter will be:

$$\sigma(\mathbf{z}) = \exp\left(\beta_0 + \beta_1 z_1 + \beta_2 z_2\right)$$

where the β_0 parameter corresponds to the effect of wood, β_1 corresponds to the additional effect of grassland, above that of wood alone, and β_2 corresponds to the additional effect of scrub, above that of wood alone. z_1 and z_2 are indicator variables (i.e., they take values 0 or 1 to indicate which habitat each observation corresponds to).

Non-factor covariates must be numeric, and are treated just as in standard linear regression. For example, imagine a non-factor covariate "Beaufort" (Beaufort is a measure of wind speed). The model for the scale parameter will be:

$$\sigma(\mathbf{z}) = \exp\left(\beta_0 + \beta_1 z_1\right)$$

where β_0 is the intercept (i.e., the effect of Beaufort when Beaufort is 0) and β_1 is the slope.

Some covariates could be either factor or non-factor. In the example above, we could specify "Beaufort" as a factor covariate, with one level for each Beaufort level. Specifying Beaufort as a non-factor covariate means that we have to fit fewer parameters, but also assumes an exponential relationship between the scale parameter and Beaufort level. Which is best depends on the data – we could use a model selection criterion (such as AIC) to help choose.

More complex models are constructed along similar lines. For example, a model containing habitat as a factor covariate and Beaufort as non-factor covariate will be:

$$\sigma(\mathbf{z}) = \exp(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3)$$

where the β_0 parameter corresponds to the effect of wood and the intercept of Beaufort, β_1 corresponds to the additional effect of grassland, β_2 corresponds to the additional effect of scrub, and β_3 is the slope of the Beaufort covariate.

Interactions between covariates can be modeled by creating a new field in the Distance data sheet that contains one covariate multiplied by the other. We hope to extend the MCDS engine to allow easier modeling of interactions in a future version.

Estimating the Detection Function at Multiple Levels

In some cases, it is useful to fit the detection function at one level, but use the fitted model to estimate f(0) (or h(0) for point transects) and probability of detection at a lower level. For example, imagine you want to estimate density by stratum, but don't have enough data to fit a separate detection function for each stratum. One solution is to fit a global model for the detection function, but using stratum (or lower) level covariates. You can then use the fitted model to estimate a separate average f(0) (or h(0)) in each stratum, using the covariates that apply to that stratum. In this way, both the stratum and global estimates of density will hopefully be less biased than if you had used the global detection function estimate.

The same argument applies if you want to estimate density by sample (e.g., transect) – for example as input to another analysis such as a spatial or habitat modelling exercise. One rarely has enough data to fit a separate detection function to each transect. However, by including transect (and lower) level covariates in a global or stratum level detection function model, you can then estimate average f(0) at the transect level using the observations, and their corresponding covariates, that apply to that transect.

To fit the detection function at one level, but estimate at a lower level, you tick both levels in the Estimate tab of the Model Definition Properties dialog (see Estimate Tab - CDS and MCDS in the Program Reference for further details).

Variance estimation

A restriction if you fit detection function at one level and estimate at a lower level is that the variance of density at the higher level must be estimated using the bootstrap. This is because Distance estimates density at the higher level by combining the density estimates from the lower level. The lower level density estimates were calculated using the same detection function model, and so are not independent, so the appropriate formula for variance at the higher level is complex – which is why we use the bootstrap instead.

Another issue is determining appropriate degrees of freedom for estimating confidence limits at the lower level. Currently, we divide the degrees of freedom from the higher level according to the number of observations for each estimate at the lower level. For example, consider a global detection function model that has 96 degrees of freedom, used to estimate f(0) at the stratum level for two strata, one with 76 and and the other with 24 observations. The degrees of freedom for the stratum level estimates will be 96*(76/100)=72.96 and 96*(24/100)=23.04 respectively.

Cluster Size as a Covariate

When objects occur in clusters, the detection function often shows "size-bias" – that is large clusters are more likely to be seen further from the line than small clusters. In conventional distance sampling, there are several methods for

dealing with this (Buckland et al. 2001), including predicting expected cluster size at zero distance using a regression of cluster size against distance (or against probability of detection). The expected cluster size is then used when converting the estimated density of clusters into density of individuals. In Distance, these options are accessible in the **Cluster Size** tab of the **Model Definition Properties** window.

In MCDS, there is another alternative: cluster size can be included in the detection function model as a covariate. In this case, the size bias will be allowed for in the detection function model. Density of individuals can be obtained directly from the Horvitz-Thompson-like estimator used in the MCDS engine (Marques and Buckland 2004), so the options in the **Cluster Size** tab become obsolete (indeed, this tab is greyed out when you select cluster size as a covariate).

When cluster size is a covariate, several options also change in the **Model Definition Properties** window **Estimate** tab. The changes are outlined below; see also the Estimate Tab - CDS and MCDS page of the Program Reference.

Stratification and post-stratification

A restriction when you select cluster size as a covariate is that stratification and post-stratification is no longer possible – see Stratification and Post-stratification in MCDS for more on this).

Variance estimation

A second restriction when you select cluster size as a covariate is that the variance of the estimate of density of individuals is not estimated analytically, but is instead obtained using the bootstrap. We also use the bootstrap to obtain a variance for the estimated expected cluster size. Analytic formulae for these variances are given by Marques and Buckland (2004) but are not currently implemented in the software.

Scaling of Distances for Adjustment Terms

In the key function + adjustment term formulation of Buckland et al. (2001, Section 2.4), the detection function is defined using a parametric key function which is then "adjusted" using a series expansion to give a flexible model. The series expansion (or adjustment terms) are one of three forms:

Series expansion	Form[1]
Cosine	$\sum_{j=2}^{m} a_j \cos(j\pi y_s)$
Simple polynomial	$\sum_{j=2}^{m} a_j y_s^{2j}$
Hermite polynomial	$\sum_{j=2}^{m} a_j H_{2j}(y_s)$

([1]Note that when a uniform key function is used in CDS, the summation is from j=1 to m)

Here, y_s is the perpendicular distance, standardized to avoid numerical problems. In Buckland et al. (2001), the standardization $y_s=y/w$ is assumed, where w is the right truncation distance. However, in the case of the multiple covariate engine, this means that the adjustment of the detection function is independent of the covariate values, so the shape of the detection function will change at different values of the covariate. This may be appropriate in some cases, for example

when search effort was conducted in such a way that the probability of detection is consistently higher at a given distance than the model without adjustments would predict. Nevertheless, to enable models with a truly constant shape to be fit in the presence of adjustment terms, Distance also offers the ability to scale the distance by σ, the scale parameter of the half-normal or hazard rate key functions (see Defining MCDS Models on page 3 of this chapter for more on the scale parameter). Since σ is a function of the covariates, this means that the scaling will be different at each covariate level, and the shape of the function will be preserved.

This option is set in the Model Definition, under the **Detection function** | **Adjustment terms** tab. For more information, see the Program Reference pages on the Model Definition Properties Dialog.

MCDS Analysis Guidelines

This section gives specific advice on doing analyses with the MCDS engine. For more general guidelines for approaching analyses in Distance, see Chapter 7 - Analysis in Distance. See also Marques and Buckland (2004) and Marques et al. (2007).

Choosing Covariates to Include in the Model

When specifying one or more covariates to be included in the model for the detection function, care must be taken to ensure that they do not violate any inherent assumptions of the method. Of primary importance are the following:

Covariates should be independent of (perpendicular or radial) distance.

A fundamental assumption of the theory behind the MCDS engine is that the distribution of distances is independent of the distribution of the covariates. If this assumption is violated, results from the MCDS engine are no longer valid or reliable. Imagine you carried out line transect surveys while walking along a path, and recorded vegetation type for each detected object as either low shrubs or tall shrubs. Imagine also that the vegetation near or around the path consisted primarily of low shrubs, whereas farther away from the path there were mostly tall shrubs. Clearly in this example, vegetation type will depend on perpendicular distance, as low shrubs will only occur at smaller distances, whereas taller shrubs will only occur at greater distances. Hence the inclusion of vegetation type as a covariate in this case will not be appropriate.

Avoid using covariates that are strongly correlated.

The use of highly correlated covariates may result in poor estimates of the detection function, or highly correlated parameter estimates, or both. If you have two covariates that are strongly correlated, and would like to include them in the model, a better approach would be to include one covariate at a time, fitting the model separately to each of them, and selecting the covariate which gives the best model fit.

Covariates should only affect the scale parameter.

It is possible to have factor covariates with levels that exhibit different shapes for the detection function – for example, the distribution of distances for one of the levels may be almost uniform, whereas for the other(s) it may follow a half-normal or hazard-rate shape. In such cases you will almost invariably end up with a very poor model fit. The use of graphic tools during exploratory data analysis (e.g. examination of histograms of the distances stratified by factor levels), prior to the actual modeling of the data, is useful for identifying this type of problem.

Specifying the Model

Fitting the detection function with multiple covariates is significantly harder, computationally, than in the case where there is only one covariate. This has several consequences:

- the analysis engine takes longer to run
- the algorithm will fail to converge more often

Because of this, it is important to be careful and thoughtful when setting up and running MCDS analyses. Here are some recommendations:

- Rather than including numerous covariates at once, start by including one covariate at a time, and selecting the covariate that gives the best model fit/lowest AIC/AICc/BIC. You can then carry out forward stepwise selection by adding one additional covariate at a time, while retaining the one(s) already selected, until there is no decrease in the AIC/AICc/BIC value (depending on the criteria you are using).
- Bear in mind that factor covariates are usually harder to fit than non-factor covariates, especially as the number of factor levels increases (see Factor and Non-factor Covariates in MCDS earlier in this Chapter for information on factor covariates). If you encounter problems while trying to fit a factor covariate, try reducing (condensing) its number of levels, if possible.
- Avoid using the feature that allows automatic selection of adjustment terms – at least to start with. Instead, start by using a model with no adjustments, and if this converges try one with one adjustment. Gradually work up to more adjustments if required.

The default for the MCDS engine's **Model Definition Properties** is for no adjustment terms. This is different from the default for the CDS engine.

One advantage of automatic forward or sequential selection is that the parameter estimates from the previous fit are used as starting values for the next model. So, for example, the parameter estimates for the fit with 0 adjustment terms would be used as starting values when trying 1 adjustment term. This helps the algorithm to converge. So, it is sometimes helpful to use automated selection, but to set the maximum number of adjustment terms to a low value, such as 2, so that too many terms are not tried. Alternatively, you can set starting values manually (see below).

- Use the option in **Model Definition Properties**, **Misc.**, **Report results for each iteration of detection function fitting routine**. This will give you some clues about whether the parameter estimates have converged.

This option is on by default in the MCDS engine.

- Another check for convergence is to compare the fitted likelihood with that from a CDS analysis with the same key function and adjustment terms. Assuming that the CDS analysis converged, the MCDS analysis should always have a likelihood that is as high, or higher. (This is because the CDS analysis contains a subset of the covariates in the MCDS analysis, so it must fit the data as well or worse.)

- Convergence is often very sensitive to the starting values used. You can set starting values manually using **Model Definition Properties**, **Detection Function**, **Adjustment Terms**, **Manually select starting values**.

Truncation for MCDS Analyses

Because MCDS methods are based on a Horvitz-Thompson like estimator of abundance, in which the inclusion probabilities are estimated (Marques and Buckland 2001, 2004), abundance estimates can be sensitive to errors in the estimated probabilities. The sensitivity is greater for smaller estimated probabilities. As a rough guide, we recommend that the method not be used if more than say 5% of the estimated probabilities of detection of observed objects, given that they were within the strip and with given covariate values, are less than 0.2, or if any are less than 0.1. If these conditions are violated, the truncation distance can be reduced. This causes some loss of precision relative to standard distance sampling without covariates.

For MCDS analyses, a summary of the proportion of estimated detection probabilities in bands 0.0-0.1 through to 0.9-1.0 is given at the bottom of the Detection Fct | Parameter Estimates page of results.

The probabilities discussed here are a function of the observed covariate values, but not of distance from the line or point; distance has been integrated out, to give more robust estimation of abundance (Marques and Buckland 2001, 2004).

Output from MCDS Analyses

The MCDS engine produces very similar output to the CDS engine (see the section Output from CDS Analyses in Chapter 8). Differences are in the Results Details listing, as outlined below.

MCDS Results Details Listing

MCDS output differs from CDS output only in the detection function pages (those that start with **Detection Fct**). The following detection function pages are output:

- **Model fitting.** Similar to CDS output, except that the results of each iteration of the fitting engine are reported by default. The model formulae and parameter indexing (A(1), A(2), etc) are also slighly more complicated, as there are now covariate parameters, but this should be self-explanatory. (For more about model formulae and parameter indexing, see About CDS Detection Function Formulae in Chapter 8.)
-

Parameter estimates. Similar to CDS output, except for the addition of a table summarizing the distribution of the estimated probability of detection, given the covariate values. This is useful for diagnosing possible problems with the estimates due to low estimated detection probabilities – see Truncation for MCDS Analyses in this Chapter.
- **Plot: Qq-plot.** Similar to CDS output. For more about CDS Qq-plots, see Chapter 8.

- **K-S GOF Test**. Similar to CDS output. For more about these CDS Goodness of Fit Tests, see Chapter 8.
- **Plot: Detection probability.** These plots can be used in the same way as for the CDS engine – for comparing the estimated function with the histograms of counts (or scaled counts for point transects). However, because covariates affect the detection function, there is no one single detection function to display. Instead, the plotted functions are the average detection function, conditional on the observed covariates. The histograms show observed frequencies at given distances, pooled over all covariates. For more information, see Chi-square GOF tests and related plots in this Chapter.
- **Plot: Pdf.** (Point transects only) See above.
- **Chi-sq GOF test.** See above.
- **FC*x*.** When there are factor covariates, you get a set of diagnostic plots for each factor combination, using only the data for that combination. For example, if there are two factor covariates, one with 3 levels and the other with 4, there will be 3x4=12 possible factor combinations (although some may not occur in the dataset, so will not be shown).
  - **Plot: Det Prb** – the detection probability plot for the given factor combination. If there are any non-factor covariates, then the plotted function will be the average detection probability, conditional on the observed non-factor covariates, and the histograms will show observed frequencies pooled over the non-factor covariates.
  - **Plot: Pdf.** (Point transects only) See above.
  - **Plot: Examp Det Funcs.** If there are any non-factor covariates, then the above plots show the detection function (or pdf) averaged over the observed covariate levels. Depending on the observed covariates, the shape of these functions can be quite different from the detection function given fixed covariate values. So, for each non-factor covariate, the MCDS engine outputs a plot showing 3 example detection functions. These are evaluated at the 25th, 50th and 75th quartiles of the covariate. If there is more than one non-factor covariate, the values of the other covariates are fixed at the 50th percentile.

Currently the example detection functions are evaluated at the observed quartiles of non-factor covariates. An alternative would be to estimate the population quartiles, using a Horvitz-Thompson-like estimator. We may implement this in a future version of Distance.

- **Plot: Examp Det Funcs.** If there are no factor-covariates, then there are no factor combination plots to display. For each non-factor covariate, the engine outputs a plot showing 3 example detection functions, as described above.

For information about how to export the results text or plots into another program, see Exporting MCDS Results.

Chi-square GOF Tests and Related Plots

In the CDS engine, the detection function depends on distance alone, and this function is displayed in the detection probability plots. By contrast, in the MCDS engine, probability of detection depends on other covariates, so there are many possible detection functions, depending on the covariate levels. Hence detection probability plots show the average detection function, conditional on the observed covariates. Similarly, for point transects, the probability density function (pdf) shown is the average pdf conditional on the observed covariates.

For example, the average conditional pdf at distance *j* is calculated as:

$$f(j \mid \mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n}\frac{f(j,\mathbf{z}_i)}{\int_{x=0}^{w} f(x,\mathbf{z}_i)dx}$$

where $f(x,z)$ is the joint density function, n is the number of observations, and w is the truncation distance. (Assuming no left truncation, otherwise x=0 in the integral is replaced by x=left truncation distance.)

Expected frequencies for the chi-square GOF test are calculated similarly. For the *j*th bin, with cutpoints c_{j1} to c_{j2}, expected frequency is:

$$\hat{E}(n_j) = \sum_{i=1}^{n}\frac{\int_{c_{j1}}^{c_{j2}} f(x,\mathbf{z}_i)dx}{\int_{0}^{w} f(x,\mathbf{z}_i)dx}$$

As with the CDS engine, the observed frequencies correspond to the total number of observations which fall within each bin.

Exporting MCDS Results

The methods of exporting results from MCDS analyses to other programs are the same as those for CDS analyses, as documented in Exporting CDS Results in Chapter 8.

In Exporting CDS Results from Analysis Details Results we show how to reproduce detection function, pdf and qq-plots in R. MCDS results contain one addition type of plot: the example detection functions produced with non-factor covariates. These can be reproduced using the following code:

```
#this reads in the file just created
forplot<-read.table(file="plot.txt", header=T, sep="\t", dec=".")
#note, depending on your language, dec might be "," rather than "."
#this plots the detection function or pdf (if point transects)
plot(forplot$C1, forplot$C2, type="l", ylim=c(0,1),
     xlab="Distance", ylab="Detection probability")
#Define labels as you wish
#this adds in other two lines
lines(forplot$C3, forplot$C4, lty=2)
lines(forplot$C5, forplot$C6, lty=3)
```

Miscellaneous MCDS Analysis Topics

Missing Data in MCDS Analysis

For information about how to deal with missing distance and cluster size data, see Missing Data in CDS Analysis in Chapter 8.

If you have missing covariate data for some observations, then this can be treated in a similar way to missing distances (see above).

If you run an analysis with data that includes missing covariates, the MCDS engine will issue a warning and exclude each observation with a missing covariate value.

Stratification and Post-stratification in MCDS

In most cases, the stratification and post-stratification options for MCDS are the same as those for CDS analyses – see Stratification and Post-stratification in Chapter 8 on CDS. An exception is when cluster size is a covariate in the detection function (see Cluster Size as a Covariate). In this case, Distance currently does not allow stratification or post-stratification. This is not an inherent limitation of the theory, rather a limitation of the current analysis engine software.

Running MCDS Analyses from Outside Distance

The MCDS engine is implemented as a stand-alone FORTRAN program, MCDS.exe. This program is called behind the scenes by Distance when you press the **Run** button on the Analysis Details Inputs tab.

Some users may wish to run the engine from outside the Distance interface – either from the Windows command line or from another program. For example, you may want to automate the running of analyses for simulations, or you may want to perform a more complicated bootstrap than Distance allows.

Full documentation for running MCDS.exe is provided in the Appendix - MCDS Engine Reference.

Analysis of Double Observer Data with the MCDS Engine

Double observer data comes from surveys where two (semi-) independent observer teams perform a distance sampling survey and duplicate detections are identified. The method for setting up double observer data in Distance is outlined in the topic Setting up a Project for MRDS Analysis in Chapter 10 of the Users Guide.

There are two ways to achieve an MCDS (or CDS) analysis of double observer data in Distance:

- Analyze the data using the MRDS engine, with **Detection Function | Method** in the model definition set equal to **ds – single observer**. Using this approach, duplicate observations are automatically removed. For more information, see Single Observer Configuration in the MRDS Engine in Chapter 10 of the Users Guide.
- Analyze the data using the MCDS engine. Using this approach, it is necessary to use the data filter to specify either observer 1 or observer 2 data or both should be used.

Here, we focus on the second option.

To analyze double-observer data using the MCDS engine, you set up an analysis where the analysis engine option in the associated model definition is MCDS.

However, it is also necessary to set up a data filter specifically to achieve the desired analysis. This is because double-observer data is entered into distance with two records for each detected object in the observation layer. So, an analysis that ignores this will have two records for each object, and so more data than there should be. Double observer data has two fields that can be used in conjunction with the data selection options in the Data Filter to achieve the desired analysis. The two fields are observer (which is 1 for the first observer and 2 for the second) and detected (which is 1 if the animal was detected by that observer and 0 if not; note that the actual field names may be different from this). For more details on these, see Setting up a Project for MRDS Analysis in Chapter 10 of the Users Guide.

Three types of analysis can be envisaged:

12. An analysis of only objects detected by observer 1. To achieve this, set up a Data Filter with data selection at the Observation layer, and the data selection criterion
    `(observer=1) and (detected=1)`
    Note that "observer" and "detected" are the default field names for these fields, but the actual field names in your project may be different if you set up the double-observer project manually (rather than via the Setup Project Wizard).
13. An analysis of only objects detected by observer 2. To achieve this, you want the data selection criterion
    `(observer=2) and (detected=1)`
14. An analysis of all objects detected, regardless of which observer detected them. To achieve this, you want the data selection criterion
    `(observer=1)`
    or you could equally use
    `(observer=2)`
    This assumes that the distance and other covariate data are the same for both observers, and also that there are no records in the dataset where detected=0 for both observer 1 and observer 2 (i.e., no-one saw it!). If both of these criteria are met, then choosing either observer=1 or observer=2 will result in half of the data being discarded, so that the number of observations going into the MCDS engine will be the number of unique objects detected, which is what you want.

Chapter 10 - Mark Recapture Distance Sampling

Introduction to Mark Recapture Distance Sampling

This entire chapter is for advanced users only!

Mark Recapture Distance Sampling (MRDS) refers to the analysis of double-observer distance sampling data, as described by Laake and Borchers (2004). Double-observer methods allow estimation of the probability of detection at zero distance, g(0), in contrast to conventional methods, where this probability is assumed to be 1.

The MRDS engine in Distance implements these methods, and this chapter describes how to use the MRDS engine. We do not describe the methods at all – instead we assume the reader is familiar with Laake and Borchers (2004). Please note that double-observer methods are an advanced technique that should only be considered once you are familiar with conventional distance sampling.

The MRDS engine currently has the following capabilities. We expect to extend these in future versions.

- Estimation for independent observer and trial configurations (see Introduction to MRDS Models; removal configuration to follow in a future version)
- Point and full independence models (in the forms of Models 1 and 3 from Table 6.1 of Laake and Borchers 2004)
- Estimation for single observer CDS and MCDS surveys (see Single Observer Configuration in the MRDS Engine; support for mixed single and double platform surveys may follow).
- Line transect surveys (point transects to follow)
- Perpendicular distance (radial distance and angle to follow)
- Data where the object of interest is an individual or cluster (see Clusters of Objects in MRDS)
- Right truncation of distances, and left truncation for single observer configuration (left truncation for double observer configuration to follow)
- Exact and interval data (manual specification of cutpoints only; more complex interval data, where intervals vary by observation to

follow, and even more complex non-continuous intervals are planned). Note that you can't do left or right truncation when the data are in intervals.

- Up to one level of geographic stratification (see Stratification and Post-stratification in MRDS; extension to multiple levels of stratification and other types of stratification may follow)
- Three analytic methods of variance estimation (see Variance Estimation in MRDS; bootstrap may follow)
- Estimation of density/abundance using a detection function fitted in a previous analysis – allows different subsets of the data to be used for encounter rate and detection function fitting (see Using a Previously Fitted Detection Function to Estimate Density in MRDS).

The MRDS engine is implemented as a library in the free statistical software R. When you run an MRDS analysis from Distance, Distance creates a sequence of R commands, calls the R software, waits for the results and then reads them back in. Therefore, before you can use the MRDS engine, you must first ensure that you have R correctly installed and configured. For more on this, see R Statistical Software in Chapter 7 of the Users Guide.

To analyze double-observer data in Distance, you then need to set up the project appropriately and include data in the correct format – see Setting up a Project for MRDS Analysis. You must next create one or more model definitions that specifies to use the MRDS analysis engine, and associate these model definitions with analyses, which you can then run. For more about the basics of setting up analyses, see Chapter 7 - Analysis in Distance. More details of the various models available in the MRDS engine are given in Defining MRDS Models, and a detailed description of the options available in the Model Definition Properties pages for this engine is given in the Program Reference pages Model Definition Properties - MRDS.

If you are new to Distance, we strongly recommend you familiarize yourself with the CDS analysis engine, for example by working through Chapter 3 - Getting Started before trying to analyze MRDS data.

The Golftees sample project is an example of how to set up a double-observer study and specify analyses – see Sample Projects.

In this chapter we also provide some analysis guidelines, give a list of the output the engine can produce and cover various miscellaneous topics such as how to deal with interval data, stratification, etc.

If you are familiar with the R software, you can run the MRDS engine directly from within R, bypassing the Distance interface altogether. For more information, see Running the MRDS Analysis Engine from Outside Distance.

There is an extensive online help that accompanies the MRDS R library. This contains more details about the methods and options available, and will be of use to users familiar with R. To open these help pages from within Distance, choose **Help | Online manuals | MRDS Engine R Help (html)**. The main functions to look at are `ddf()` and `dht()`.

Setting up a Project for MRDS Analysis

The easiest way to set up a new project for an MRDS analysis is using the Setup Project Wizard.

- In Step 1, under **I want to:**, select **Analyze a survey that has been completed**.
- In Step 3, under **Observer configuration**, select **Double observer**.
- Follow through the rest of the wizard as usual.

Distance then creates the appropriate data fields for double observer data, and you can then import your data using the Import Data Wizard. For more about the data requirements, see Setting up your Data for MRDS Analysis.

Alternatively, you can create the appropriate fields by hand, and manually create a new survey object with the appropriate observer configuration and data files. For more about survey objects, see Working with Surveys During Analysis in Chapter 7.

Setting up Your Data for MRDS Analysis

When you use the select a double observer configuration in the Setup Project Wizard, Distance creates three extra data fields in the Observation layer (in addition to the appropriate distance and cluster size fields). These are:

- **object** – this must contain a unique integer for each object detected by either observer. There should be two records in the observation layer for each object seen – one for observer 1 and the other for observer 2.
- **observer** – tells Distance which observer the record refers to: either 1 or 2.
- **detected** – tells Distance whether the object was detected by that observer or not: 1 for detected, 0 for not detected.

The fields do not have to have those names – if you set up the project manually, you can call them what you like. However, in the Data Fields tab of the survey object for double observer configurations, you have to tell Distance which fields play the object, observer and detected roles.

An example of double observer data, from the Golftees sample project, is shown below:

Contents of Observation layer 'Observation' and all fields from higher layers

	Observation						
ID	Perp distance	Cluster size	object	observer	detected	sex	exposure
ID	Decimal	Decimal	Integer	Integer	Integer	Integer	Integer
n/a	m	[None]	[None]	[None]	[None]	[None]	[None]
Int	Int	Int	Int	Int	Int	Int	Int
1	2.68	2	1	1	1	1	1
2	2.68	2	1	2	0	1	1
3	3.33	2	2	1	1	1	0
4	3.33	2	2	2	0	1	0
5	0.34	1	3	1	1	0	0
6	0.34	1	3	2	0	0	0
7	2.53	2	4	1	1	1	1
8	2.53	2	4	2	0	1	1
9	1.46	2	5	1	1	1	0
10	1.46	2	5	2	0	1	0

Part of the Observation layer from the Golftees double observer example project

If you open the Golftees sample project you will notice that:

- The object number goes up sequentially in this example (1, 2, 3, …) – in general the object number should be unique but it doesn't need to be sequential.

- In this example the two records for each object come one after the other (e.g., lines 1 and 2 are Observer 1 and Observer 2 for object 1) – in general they don't have to so long as there are two lines for each object – one with Observer 1 and one with Observer 2. For example, you might like to structure your data with all records for observer 1 first and then all records for observer 2
- The detected field indicates whether an observer saw the object or not. For example, object 1 was seen by observer 1 but not by observer 2.
- In this example, the distance field (Perp distance) contains the same distance for both observers, regardless of whether the observer saw the object or not. In general you should always put the same distance for both observers – a version that can deal with measurement error and so allow different distances is planned.
- There are some additional covariates in the observation layer in this example: sex, exposure and Cluster size. In general, covariates can be placed in any of the data layers – although the rules for referring to the covariates differ between the layers – for more on this, see Defining MRDS Models.
- In this example, there is only one transect, and so there are no transects on which no objects were seen. In general there may well be transects with no objects. On these transects you should not enter any observations (just as with the CDS engine – see the Example Data Sheet picture on the Data Fields page in Chapter 5 for an example, and see Introduction to Data Import for how to import such data).

Defining MRDS Models

Introduction to MRDS Models

Specifying a detection function in MRDS requires specifying the form of up to two functions (Laake and Borchers, 2004, section 6.3.2.3). The first is the unconditional detection function $g_{\cdot}(y,\underline{z})$ – the probability of one or more observer detecting the object, given it's distance and covariate values. The second is the conditional detection function $p_{j|3-j}(y,\underline{z})$ – the probability of observer *j* detecting the object, given that the other observer (observer 3-*j*) has detected it and also given its distance and covariate values. We refer to the former as the distance sampling or DS model and the second as the mark recapture or MR model. Which of these needs to be specified depends on the fitting method chosen, as follows:

Fitting method	Configuration	Independence	DS model	MR model
ds	single[1]	-	yes	no
io	independent	point	yes	yes
io.fi	independent	full	no	yes
trial	trial	point	yes	yes
trial.fi	trial	full	no	yes
removal[2]	removal	point	yes	yes
removal.fi[2]	removal	full	no	yes

[1]*For more on this, see the topic Single Observer Configuration in the MRDS Engine later in this chapter.*
[2]*These methods are not implemented in the current version of Distance*

The fitting method is chosen on the **Detection Function** | **Method** page of the Model Definition Properties.

DS and MR Models

The form of the two DS and MR models are different. The implementation here corresponds with models 1 and 3 of Table 6.1 from Laake and Borchers (2004) – see also that reference for more information.

DS Model

The DS model is of the same form as the models used in the MCDS engine, except that currently only the key function part of the model is implemented, with no adjustment terms. The two key functions implemented are half-normal and hazard rate:

Half-normal key function, $\exp\left\{-x^2/2\sigma(\mathbf{z})^2\right\}$

Hazard-rate key function, $1-\exp\left\{-\left[x/\sigma(\mathbf{z})\right]^{-b}\right\}$

The scale parameter, $\sigma(\underline{z})$, is modeled as an exponential function of the covariates:

$$\sigma(\mathbf{z}) = \exp\left(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots \beta_q z_q\right)$$

The key function and covariates to use are specified on the **Detection Function** | **DS Model** page of the Model Definition Properties. One can also choose no additional covariates, which corresponds with a CDS model. Note that distance (x in the above formulae) is automatically a part of this model.

If you're comparing DS model results with those obtained from the MCDS engine, you should note that in the MCDS engine, the scale parameter is modeled as $\sigma(\mathbf{z}) = \beta_0 + \exp(\beta_1 z_1 + \beta_2 z_2 + \ldots \beta_q z_q)$. Therefore to compare the two values of β_0, you need to exponentiate the β_0 from the MRDS Engine's DS model.

MR Model

The MR model is currently implemented as a logistic model – i.e. it is of the form:

$$p_{j|3-j}(y,\underline{z}) = \frac{\exp(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots \beta_q z_q)}{1+\exp(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots \beta_q z_q)}$$

The covariates to use are specified on the **Detection Function** | **MR Model** page of the Model Definition Properties. Note that these covariates need not be the same as those chosen for the DS model, if there is a DS model with the chosen fitting method. Note also that unlike for DS models, distance is not automatically a part of the model – if you want to include distance as a covariate in an MR model you must explicitly include it in the model formula.

Specifying DS and MR Model Formulae

For both DS and MR models, one must specify formulae that tell Distance which covariates to include in the model (the general form of the two models are given in the previous section, DS and MR Models). These formulae consist of a series

of terms joined by operators such as "+". The terms represent covariates and the operators tell Distance how the covariates relate to one another.

For example, the MR formula

```
distance + sex + exposure
```

means include the data from the fields distance, sex and exposure as covariates.

To understand how to specify formulae, we need to understand (1) how to translate a field in the Distance database into a covariate to specify and (2) what are the possible operators and how do they work. We also need to understand the difference between factor and non-factor covariates and how to specify which is which. These are covered in the following sections.

To see some examples of model formulae, have a look at the Golftees example project.

Translating Distance Fields into DS and MR Covariates

Unfortunately, there is not a 1:1 correspondence between the field names in the Distance project database and the covariate names you can use to specify DS and MR model formulae. This is due to some limitations of the R statistical language that the MRDS engine is implemented in, and also in the way the MRDS engine is written.

Field Translation Made Simple

The easiest way to work out how to write your formulae is to run a simple MRDS analysis in Distance, perhaps with a trial.fi fitting method and "1" as the MR model (i.e., and intercept-only model). Then, look in the log tab for something like the following, which lists the field name and translated covariate name for all fields in the database:

```
The following fields will be written to the detection function data
file, and can be used in detection function model formulae. Note
that you should use the new names, not the original field names in
formulae, and that formulae names are case sensitive.
  Format: [layer name].[field name] AS new name
  [Observation].[Perp distance] AS distance
  [Observation].[Cluster size] AS size
  [Observation].[object] AS object
  [Observation].[observer] AS observer
  [Observation].[detected] AS detected
  [Observation].[sex] AS sex
  [Observation].[exposure] AS exposure
  [Line transect].[Label] AS label
  [Line transect].[Line length] AS line.length
  [Region].[Label] AS stratum.label
  [Region].[Area] AS area
  [Study area].[Label] AS global.label
```

For example, if you then wanted to specify a formula with the label field from the region layer (i.e., [Region].[Label]) and observer from the Observation layer (i.e., [Observation].[observer]) as covariates, you would write the formula as:

```
stratum.label + observer
```

Field Translation in Detail

If you want to try to predict what translations Distance will do, here are the rules it applies:

- You can specify covariates using fields in any data layer. However, if the field name occurs in more than one data layer, and so is not unique, then Distance has to use some method to determine which layer you are referring to. So it does the following. It starts at the lowest layer and stores the field names. If it encounters the same name in a higher layer

it renames it by adding the layer type and a dot in front of the field name. For example, by default, there is a field called "Label" in the sample, stratum and global layers. The sample layer is the lowest, so the formula name for this is `label` (note it is lower case as all formula names are changed to lower case – see below). The stratum and global layers are higher, so the formula names for the Label fields in those layers are `stratum.label` and `global.label`.

- To comply with R object naming rules, and make life simpler:
    - spaces and anything else that isn't a letter or number are replaced by dots – e.g., "type of habitat" becomes `type.of.habitat`.
    - all letters become lower case – e.g., "Type of Habitat" becomes `type.of.habitat`.
    - fields that don't start with a letter get an "X" appended - e.g., "1 covariate" becomes `X1.covariate`.
- Fields with certain roles are always translated to a fixed name, as follows.

Role	Translated to
Perpendicular distance	distance
Cluster size	size
Object	object
Observer	observer
Detected	detected

As an example, if you have a field "Cluster size" which is defined as the having the role of cluster size in the survey definition, then the name you should specify in MR or DS formulae is `size`.

DS and MR Model Formulae Operators

- **+** means the two covariates on either side of the operator are both in the model as main effects – e.g., a + b means include terms for covariate a and covariate b
- **:** means the interaction of the two covariates on either side of the operator – e.g., a:b means the interaction of a and b
- ***** means the two main effects plus all interactions – e.g., a*b is the same as a + b + a:b. More than two covariates can be included – e.g., a*b*c is the same as a + b + c + a:b + a:c + b:c + a:b:c
- **^** indicates crossing to a specified degree – e.g., (a+b+c)^2 is the same as (a+b+c)*(a+b+c) which in turn expands to a formula containing the main effects for a, b and c together with their second-order interactions
- **%in%** indicates that the terms on its left are nested within those on the right. For example a + b %in% a expands to the formula a + a:b
- **-** removes the specified terms, so that (a+b+c)^2 -a:b is identical to a + b + c + b:c + a:c. It can also used to remove the intercept term: x - 1 is a line through the origin. (A model with no intercept can be also specified as y ~ x + 0 or 0 + x)

More About DS and MR Model Formulae

- An intercept term is included in formulae by default. To remove it you can use "- 1" - for example `sex - 1`, while to specify an intercept-only formula, you use `1` alone.

Aside! As an example of the use of the use of an intercept-only formula, one way to specify a Peterson model is using a trial configuration where there is full independence and an intercept-only MR model.

- While formulae usually involve just variable and factor names, they can also involve arithmetic expressions. The formula `sex + log(size)` is quite legal. Take care if you use arithmetic expressions that there isn't a field with the same name (e.g., a field in the observation layer called "log")!

Factor and Non-factor Covariates in MRDS

For an explanation of the difference between factor and non-factor covariates, see Factor and Non-factor Covariates in MCDS. In the MRDS engine, all covariates are assumed to be non-factor, unless specified otherwise. To specify a covariate as a factor, include it as part of a comma-delimited list in the **Detection Function | Factors** page of the Model Definition.

For example, if the MR model is given as `distance * sex * exposure`, and the Factors page specifies the following: `sex, exposure`, then the MRDS engine will assume that distance is a non-factor covariate and that sex and exposure are factors.

Tip! You can specify covariates as factors even if they are not included in a model in the current model definition. It saves time to list all the factor covariates in your first model definition, as this list will then be copied to all subsequent model definitions that you define, saving you the bother of having to type the factor list for each model definition. In the above example, we could have specified `observer, sex, exposure` as the factor list – observer is not in the current model, but if we subsequently define a model definition based on this one and include observer as a covariate we won't have to remember to include it in the list of factors as it will already be there.

MRDS Analysis Guidelines

These methods are relatively new, so we are only beginning to gain experience on effective analysis strategies. Some preliminary guidelines follow – please let us know if you can suggest some further guidelines or amendements.

- Start with some CDS and MCDS analyses (i.e. using the CDS/MCDS engines) to get a ball park on the detection function shapes, etc. – see Analysis of Double Observer Data with the MCDS Engine in Chapter 9 for some tips on doing this. You can also perform CDS and MCDS analyses using the MRDS engine – see Single Observer Configuration in the MRDS Engine later in this chapter.
- Your first MRDS analysis could be a Peterson or other simple model – helps work out what covariate names to use (see Translating Distance

fields into DS and MR covariates), and should also fit without problems.

- Build up covariates slowly. You may need to specify starting values (although this option isn't available currently in the interface) look at the iteration history (**Detection Function** | **Control** page of Model Definition), etc. to work around any convergence problems.
- If you experience problems, check Problems with the MRDS Engine in the Troubleshooting chapter, and also check the Program Distance Web Site for the latest list of known problems.
- This is a new analysis engine – you can expect some teething problems. Contact the program authors if you can't resolve them (see Sending Suggestions and Reporting Problems).

Output from MRDS Analyses

The MRDS engine produces the following output:

- a summary of results in the **Analysis Browser**. For general information about the Analysis Browser, see the section Introduction to the Analysis Browser in Chapter 7. There are many results statistics available, and you can select which ones are shown independently for each analysis set using the Column Manager (see Column Manager Dialog in the Program Reference). An explanation of some of the columns is given in the section MRDS Analysis Browser Results.
- a detailed listing of results in the **Results** tab of the **Analysis Details** window. These are described in the following section, MRDS Results Details Listing.
- a log of the analysis, highlighting any possible problems, in the **Log** tab of the **Analysis Details** window. For information about troubleshooting problems, see Chapter 12 - Troubleshooting.
- (optionally) plots from the results details can be imported into other programs. For more about this, see the section on Exporting MRDS Results.

MRDS Results Details Listing

When an analysis has run, a great deal of information is available in the **Results** tab of the **Analysis Details** window. This information is split into pages, as follows:

- **Detection Fct/Summary.** Gives a summary of the results of the detection function modelling.

You can get more information about the fitting process using the `showit` control setting – for details, see the section on Fine-tuning an MRDS Analysis.

- **Detection Fct.** A set of pages about the detection function model fited.
  - **Plot: Qq-plot**. A quantile-quantile plot, showing the goodness-of-fit of the fitted model (exact data only). For information about what qq-plots are, see CDS Qq-plots in Chapter 8.

- **Goodness-of-fit.** Chi-squared goodness of fit tests for the DS and/or MR models (depending on the fitting method), and Kolmogorov-Smirnov and Cramér-von Mises tests (for exact data). For more about these latter tests, see CDS Goodness of Fit Tests in Chapter 8.
- **Plot: Detection Probability.** Plots of the fitted detection functions, superimposed on histograms showing the frequency of counts. The estimated probability of detection of each observation (given its covariate values and distance) is also shown. The number of plots depends on the fitting method.

The plots are stored as image files in the R directory, and so can easily be imported into other programs - see Exporting MRDS Results.

- **Density Estimates and associated quantities.**
  - Standard output consists of 3 tables: (i) a summary containing area, covered area, survey effort, number of observations and encounter rate; (ii) abundance and associated CV and confidence limits; (iii) density and associated CV and confidence limits. All are reported by stratum and pooled across strata. If objects are in clusters there are 3 tables for estimates by cluster and 3 more for estimates by individual, followed by a further table of estimated expected cluster sizes.
  - Further output is available by choosing the Extended output option in the Misc tab of the Model Definition (see Misc - MRDS in the Program Reference). This consists of: (i) variance-covariance matrix for the abundance estimates due to estimating the detection function parameters; (ii) variance-covariance matrix for the abundance estimates due to selection of samples; (iii) resulting correlation matrix of the abundance/density estimates; (iv) estimates by sample of sample effort, covered area, number of observations, estimated abundance and density.

MRDS Analysis Browser Results

Currently, a relatively limited number of statistics are available in the Analysis Browser: number of parameters, AIC, log likelihood, density and abundance with associated CVs and confidence limits. We will be expanding the list of outputs in future versions.

Exporting MRDS Results

The methods of exporting results from the Analysis Browser and results pages of the Analysis Details to other programs are the same as those for CDS analyses, as documented in Exporting CDS Results in Chapter 8. For example you can copy the results details text by choosing **Analysis Results | Copy Results to Clipboard**, and you can copy plots by choosing **Analysis Results | Copy Plot to Clipboard**. Note though that in the case of the plot, the underlying data are not copied as well (unlike for CDS plots) – just the plot picture.

There is another way to get hold of the plots produced by MRDS analyses: to directly access the image files produced when each analysis is run. Each plot that is displayed in the Results Details has a corresponding image in the project's R Folder. For more details, see Images Produced by R in Chapter 7 of the Users Guide.

Miscellaneous MRDS Analysis Topics

Interval Data in MRDS

Not implemented in the current version of this engine.

Clusters of Objects in MRDS

If the objects detected are clusters of individuals, you tell Distance this in the same way as for CDS analyses, in the Setup Project Wizard (see Survey Methods Wizard Page in the Program Reference).

Unlike CDS analysis, the Model Definition does not offer any regression methods for dealing with size bias. If you suspect size bias is a potential problem, the appropriate way to deal with it in an MRDS analysis is to include cluster size (or some transformation of cluster size) as a covariate in the detection function model(s).

The cluster size field is one of the fields with a fixed name in detection function formulae in MRDS (see Translating Distance Fields into DS and MR Covariates) – in formulae you should use the name `size` regardless of the actual field name.

Stratification and Post-stratification in MRDS

At the moment, the MRDS engine only accommodates one level of stratification, and this stratification is assumed to be geographic (i.e., the global density estimate is calculated as the mean of the stratum estimates, weighted by stratum area). More options are planned for future versions.

Variance Estimation in MRDS

The MRDS engine currently offers three analytic methods for estimating variance of the density estimate. These are:

- **Innes et al. (2002) - based on the empirical variance in estimated density between samples**. This is the default method, and the one that should generally be used. The formula has terms for variation in density between samples given the estimated detection function parameters, and for uncertainty due to estimating the detection function parameters, as follows (from Marques and Buckland et al. 2004, formula 3.27):

$$\hat{\text{var}}\left(\hat{N}_s\right)=\left(\frac{A}{2wL}\right)^2\left\{L\sum_{k=1}^{K}\frac{l_k\left(\frac{\hat{N}_{csk}}{l_k}-\frac{\hat{N}_{cs}}{L}\right)^2}{K-1}+\sum_{j=1}^{r}\sum_{m=1}^{r}\frac{\partial\hat{N}_{cs}}{\partial\hat{\theta}_j}\frac{\partial\hat{N}_{cs}}{\partial\hat{\theta}_m}H^{-1}_{jm}\left(\hat{\underline{\theta}}\right)\right\}$$

where $\hat{N}_s$ is the estimated abundance of clusters (or individuals if cluster size is always 1) in the study region (of size A), $\hat{N}_{cs}$ is the estimated abundance of clusters in the covered region (of size $2wL$, where w is truncation distance and L is the total line length), $\hat{N}_{csk}$ is the estimated abundance of clusters on the kth line (k=1…K), l_k is the

length of the k^{th} line, $\hat{\underline{\theta}}$ is a vector of length r containing the parameter estimates of the detection function model, and $H^{-1}_{jm}\left(\hat{\underline{\theta}}\right)$ is the jm^{th} element of the inverse of the Hessian matrix for $\hat{\underline{\theta}}$. (See also formula 3.35 for the variance of the number of individuals.) More details are given in Innes et al. (2002) and Marques and Buckland (2004).

- **Buckland et al. (2001) - based on the delta method, using the empirical variance in encounter rate between samples**. This method is based on the conventional distance sampling variance estimator of Buckland et al. (2001, Sections 3.6.1 and 3.6.2), extended to allow probability of detection to vary among individuals. The method assumes independence between the estimates of detection function parameters, encounter rate (and mean cluster size for variance of the estimate of abundance of individuals) – an assumption not made by the Innes et al. estimator. For that reason, the Innes et al. estimator is preferred. The formula can be written:

$$\hat{\text{var}}\left(\hat{N}_s\right)=\left(\frac{A}{2wL}\right)^2\left\{\frac{\hat{N}_{cs}L}{\left(n_s w\right)^2}\sum_{k=1}^{K}\frac{l_k\left(\frac{n_{sk}}{l_k}-\frac{n_s}{L}\right)^2}{K-1}+\sum_{j=1}^{r}\sum_{m=1}^{r}\frac{\partial\hat{N}_{cs}}{\partial\hat{\theta}_j}\frac{\partial\hat{N}_{cs}}{\partial\hat{\theta}_m}H^{-1}_{jm}\left(\hat{\underline{\theta}}\right)\right\}$$

  where n_s is the number of clusters seen, and n_{sk} is the number of clusters seen on line k.

- **Binomial variance of detection process**. This variance estimator should only be used when the entire study area is sampled (as happens sometimes, for example in simulation experiments). It only contains the term for uncertainty due to estimating the detection function parameters – i.e., it assumes no variance comes from scaling up the estimated density on the surveyed area to the whole study area. The formula is:

$$\hat{\text{var}}\left(\hat{N}_s\right)=\left(\frac{A}{2wL}\right)^2\left\{\sum_{i=1}^{n_s}\hat{f}\left(0\,|\,\underline{z}_i\right)^2-\hat{N}_{cs}+\sum_{j=1}^{r}\sum_{m=1}^{r}\frac{\partial\hat{N}_{cs}}{\partial\hat{\theta}_j}\frac{\partial\hat{N}_{cs}}{\partial\hat{\theta}_m}H^{-1}_{jm}\left(\hat{\underline{\theta}}\right)\right\}$$

  where $\hat{f}\left(0\,|\,\underline{z}_i\right)$ is the estimated pdf of observed distances given the covariate values $\underline{z}_i$, evaluated at zero distance.

The appropriate option can be selected on the Variance tab of the Model Definition Properties Dialog (see Variance - MRDS in the Program Reference).

Multipliers in MRDS Analysis

Multipliers are currently not implemented in the MRDS engine. Two main uses for multipliers are dealt with using other methods in this engine – the case when g(0)<1 is dealt with via the double-observer data, and the case where we wish to fit a detection function to some data and apply it to a subset is dealt with explicitly in the section Using a Previously Fitted Detection Function to Estimate Density in MRDS.

Model Averaging in MRDS Analysis

This is currently not implemented in the MRDS engine. For more on model averaging, see Model Averaging in CDS Analysis in Chapter 8 of the Users Guide.

Sample Definition in MRDS Analysis

This is implemented in exactly the same way as for CDS and MCDS analyses - see Sample Definition in CDS Analysis in Chapter 8 of the Users Guide for details.

Using a Previously Fitted Detection Function to Estimate Density in MRDS

Using the MRDS engine, you can fit the detection function to your survey data, and use this fitted detection function to estimate detection probability for subset of the data. This can then be combined with encounter rate information for that subset to estimate density. There are several reasons why you might want to do this. For example:

- In a multi-species study, you may want to explore fitting a combined detection function to a group of species. You can use model selection criteria such as AIC to decide whether to fit individual models by species, or a combined model with or without species (or guild) level covariates. Once the model selection is done, you are usually interested in obtaining estimates of density by species, so you would apply the global density function (if that is what was chosen) to subsets of the data for each species.
- You may have two surveys of the same species at different times but over the same range of conditions, where one has too few sightings to estimate the detection function but the other has enough. You can fit the detection function to the larger dataset (or a combined dataset) and then, given the covariate values from the smaller one, predict detection probabilities in the smaller dataset.
- You may want to fit a detection function to a relatively common species, and then apply this function to a species for which there are only a few observations (assuming you collected the same covariate information for both species).
- Rather than species, other components of the population could be the target (e.g., sex, etc).

How is this kind of analysis done in the MRDS engine?

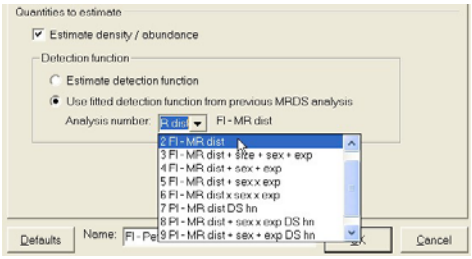
You start with the the data for which you want to model the detection function – you can select this data from the whole dataset using the Data Selection tab of the Data Filter. Then define the model definitions you want and run the corresponding analyses. Decide which analysis you want to use to provide a detection function for the new subset of data.

For these model definitions, you probably don't want to estimate density, just the detection function. You can achieve this by un-checking the option in the **Estimate** tab of the Model Definition to **Estimate density / abundance**.

Recall that the MRDS engine is actually implemented as a language in the statistical software R, and that to run an MRDS analysis, Distance writes out command and data files, then runs R in batch mode, waits for it to finish and then harvests the results from output files. By default, the R objects generated by the run are not saved by R, but are deleted. In this case, we want to overwrite that behaviour and save the fitted detection function object so that it can be used in a future analysis. We do this by selecting **Tools | Preferences | Analysis** and then under **R software**, uncheck the option Remove the new objects that are created with each run. Now run the analysis that you are going to use to

provide the detection function fit again, and the R objects will be saved (in the R object file .RData in the R Folder – see Contents of the R Folder in Chapter 7 for more on this file).

Now we'll apply this detection function to the new subset of data. Define a new data filter that selects the new subset of data. Note that it must be a subset of the data used to fit the detection function. Define a new Model Definition, and in the **Estimate** tab, check the option to **Estimate density / abundance** and also the option to **Use the fitted detection function from previous MRDS analysis**. Select the ID of the analysis you want to use. For example, if you want to use analysis 2 (called "FI - MR dist" in this case), the lower half of the **Estimate** tab will look like this:


Quantities to estimate
Estimate density / abundance
Detection function
Estimate detection function
Use fitted detection function from previous MRDS analysis
Analysis number:
2
FI - MR dist
2 FI - MR dist
3 FI - MR dist + size + sex + exp
4 FI - MR dist + sex + exp
5 FI - MR dist + sex:x exp
6 FI - MR dist x sex x exp
7 FI - MR dist DS hn
8 FI - MR dist + sex:x exp DS hn
9 FI - MR dist + sex + exp DS hn
Defaults
Name:
Cancel

Applying the fitted detection function from one analysis to another analysis

When you run the new analysis, the probability of detection for each object in the new analysis is estimated using the fitted detection function from analysis 2 (in this case).

Restricting Inference to Density or Abundance in the Covered Region in MRDS Analysis

For more on the circumstances when this may be appropriate, see Restricting Inference to Density or Abundance in the Covered Region in CDS Analysis in the CDS Chapter. It is achieved in MRDS by selecting the **Binomial variance of detection process** option in the variance tab – see the section on Variance Estimation in MRDS earlier in this chapter.

Running the MRDS Analysis Engine from Outside Distance

The MRDS engine is implemented as a library in the free statistical software R. When you run an MRDS analysis from Distance, Distance creates files containing a set of R commands and the appropriate input data, calls the R software, waits for the results and then reads them back in. For more about R, see R Statistical Software in Chapter 7 of the Users Guide.

Some users may wish to run the engine from outside the Distance interface – either from within the R GUI interface or from another program. For example, you may want to automate the running of analyses for simulations or bootstrapping.

To see the format of the input and data files produced by Distance, try running an MRDS analysis in debug mode. To set debug mode on, choose **Tools | Preferences**, **Analysis** tab, and tick **Debug Mode**. When you run analyses in debug mode, the input and data files are created, but the analysis is not run. The Log tab displays the location of the files – these are created in a directory with a name "dst" followed by up to 4 numbers, located within the Windows temporary directory. The file "in.r" contains the commands. Data are located in the files "ddf.dat.r" (used in the detection function modelling), "region.dat.r", "sample.dat.r" and "obs.dat.r" (used in estimating density given a fitted detection

function). You can use these files as templates for creating your own command and data files.

To run the analysis from within the R GUI (Graphical User Interface), you can cut and paste the commands from the file in.r. To run the analysis from another program, you can call R in batch mode – this is achieved by calling the program RCmd.exe, which is located within the /bin subdirectory of your R installation. For more details, see the R for Windows FAQ (in R, type `help.start()` and when a browser window opens, click on the FAQ for Windows port). For an example of its use, see the Log tab of any MRDS analysis you have run that was not in debug mode – you should see a line of the form:

```
Starting engine with the following command:
C:\PROGRA~1\R\rw1091\bin\Rcmd.exe BATCH C:\temp\dst90474\in.r
C:\temp\dst90474\log.r
```

Users familiar with R may wish to work inside the R GUI. The MRDS engine is contained in the library mrds. To load the library from within R GUI, type

```
library (mrds)
```

All the functions in the mrds library are documented – the main functions are `ddf()` (fits the detection function) and `dht()` (estimates abundance using the Hortvitz-Thompson-like estimator). You can open a copy of the help files from within Distance by choosing **Help | Online Manuals | MRDS R Engine Help (html)**.

The use of these libraries in operating systems other than Windows is not supported (but may well work – let us know!)

Installing an Updated Version of the MRDS Engine

The MRDS engine is implemented as a library (called "mrds") in the statistical software R. From time to time, we may issue updated versions of the library, for example in response to reported problems. These will come as an archive file **mrds.zip**. To install the new version,

- find the old version of mrds.zip in the Distance program directory.
- copy the new mrds.zip over the top of it (you may want to rename the old version first, as a backup)
- Choose **Tools** | **Preferences** | **Analysis**, and tick the option to **Re-install analysis engine library on next** run.
- Open a project containing analyses that use the MRDS engine (for example the Golftees sample project).
- Run one of these analyses.
- Check in the Log tab of the Analysis Details. You should find the line:
  ```
  package 'mrds' successfully unpacked and MD5 sums
  checked
  ```

The next topic describes how to check which version of the library is being used.

Checking Which Version of the MRDS Engine is Being Used

The MRDS engine is implemented as a library (called "mrds") in the statistical software R. From time to time, we may issue updated versions of the library, for example in response to reported problems. Therefore, before downloading a

new version or reporting a problem you may want to check which version of the library is currently in use. To do this, re-run an analysis that uses the MRDS Engine (such as one from the Golftees sample project) and look in the Log tab for the line

```
> library(mrds)
```

After it, you should see a line which looks something like the following:

```
This is mrds 1.2.7
Built: R 2.3.1; i386-pc-mingw32; 2006-08-09 17:33:03; windows
```

If you are reporting a problem you should quote both the build number (in the above case 1.2.7) and the build date and time (2006-08-09 17:33:03).

The previous topic describes how to update to a newer version of the MRDS Engine, if one is available.

When reporting results, you may want to cite the exact version (i.e., build number) of the library that used in the analysis. This is stored in the Log tab, as outlined above.

Fine-tuning an MRDS Analysis

In most cases, the default options for estimation of detection function parameters in an MRDS analysis work well. However, there are times when you need to tweak the analysis, for example by setting starting values or bounds on parameters. You may also want to get more information on the fitting process, such as parameter estimates at each iteration of the optimization. This kind of fine-tuning is specified in the **Detection function | Control** page of the model definition.

To enter options on this page, type them in as a comma-delimited list. Any noninteger numbers should have a decimal point separating the integer and fractional parts – e.g., `38.98`. You can also use engineering notation – e.g., `3.898E1`. For some options (e.g., starting values), you need to specify a vector of numbers. To do this, write them out as a comma-delimited list prefixed by `c(` and suffixed by `)` – e.g., `c(4.7, 0.1, 0.2)`. An example of a control list is:

```
showit=T, doeachint=T, lowerbounds=c(0,0,0)
```

The options you specify are exported to R without change and turned into a list object, which is why the above format is required.

The options vary slightly depending on which detection function method is being fit (ds, io, etc). More details are in the mrds R library online help, under the appropriate ddf function (`ddf.ds`, `ddf.io`, etc). However, in general, the options are as follows.

- `showit` – F (false, the default) or T (true); if true gives output at each iteration of the fit
- `doeachint` – F (false, the default) or T (true); if true uses numerical integration rather than an interpolation method during fitting
- `estimate` – T (true, the default) or F (false); if false fits the detection function model but doesn't estimate predicted probabilities.
- `refit` – T (true, the default) or F (false); if true, the algorithm tries multiple optimizations at different starting values if it doesn't converge

- `nrefits` – integer number – controls the number of refitting attempts
- `initial` – a vector if initial values for the parameters
- `lowerbounds` – a vector of lower bounds for the parameters
- `upperbounds` – a vector of upper bounds for the parameters
- `limit` – T (true, the default) or F (false); if true then restricts analysis to observations with detected=1 (this option is ignored if fitting method = ds and there is no detected field)

Single Observer Configuration in the MRDS Engine

You can analyze single observer data using the MRDS analysis engine if you set the detection function method to ds. To do this, choose **ds - single observer** from the drop-down list on the **Detection function | Method** page of the Model Definition. Then specify the detection function model you want on the **Detection function | DS Model** page.

You can expect results to differ slightly from those you can obtain using the CDS and MCDS engines, as the optimization algorithm is different. Also note that the default variance estimation method in the MRDS engine is different (see Variance Estimation in MRDS).

There is currently no great advantage to fitting single observer data in the MRDS engine, as the CDS and MCDS engines offer more features – however this may change in the future as more features are added to the MRDS engine.

You can also fit CDS and MCDS models (i.e., models that assume all animals on the trackline are detected) to double observer data by choosing the ds detection function method. When you run such an analysis, Distance will pool the data from the two observers, so that the data are the total number of unique detections.

Chapter 11 – Density Surface Modelling

Introduction to Density Surface Modelling

This entire chapter is for advanced users only!

Density surface modelling (DSM) refers to the analysis of distance sampling data, for the purposes of predicting the spatial arrangement of animals in the study region as described by Hedley and Buckland (2004).

The DSM engine in Distance implements the "count" method described in that paper, and this chapter describes how to use the DSM engine. We do not describe the methods in detail – instead we assume the reader is familiar with Hedley and Buckland (2004). Please note this is an advanced technique that should only be considered once you are familiar with conventional distance sampling.

The DSM engine currently has the following capabilities. We expect to extend these in future versions.

- Line transect surveys (strip and point transects forthcoming)
- Perpendicular distance to objects (distance to object and sighting angle not accommodated)
- Data where the object of interest is an individual or cluster (see Clusters of Objects in MRDS)
- Response variable is estimated abundance within a segment (estimated density and number counted in a segment in future implementation)
- Up to one level of geographic stratification
- Variance estimation via moving block bootstrap; with uncertainty derived from the estimation of detectability included in an ad hoc manner
- Estimation of density/abundance using a detection function fitted in a previous analysis – allows different subsets of the data to be used for encounter rate and detection function fitting (see Using a Previously Fitted Detection Function to Estimate Density in MRDS).

As with the MRDS engine, the DSM engine is implemented as a library in the free statistical software R. When you run a DSM analysis from Distance, Distance creates a sequence of R commands, calls the R software, waits for the results and then reads them back in. Therefore, before you can use the DSM engine, you must first ensure that you have R correctly installed and configured. For more on this, see R Statistical Software in Chapter 7 of the Users Guide.

To produce a density surface model in Distance, you then need to set up the project appropriately and include data in the correct format – see Setting up a Project for DSM Analysis. You must next create one or more model definitions using the MRDS analysis engine, and associate these model definitions with analyses to derive detection probabilities for each objected detected. For more about the basics of setting up analyses, see Chapter 7 - Analysis in Distance. More details of the various models available in the MRDS engine are given in Defining MRDS Models, and a detailed description of the options available in the Model Definition Properties pages for this engine is given in the Program Reference pages Model Definition Properties - MRDS. After deriving detection probabilities, then a density surface model can be fitted. In addition, you must also create a prediction grid that contains values of the covariates used for spatial prediction at a grid of locations throughout the study region, not only along the surveyed line transects. This prediction grid must be geo-referenced, and read by Distance in a particular fashion to take advantage of the spatial data contained therein. See Producing a prediction grid in GIS for further details.

If you are new to Distance, we strongly recommend you familiarize yourself with the CDS analysis engine, for example by working through Chapter 3 - Getting Started before trying to analyze DSM data.

The Dolphin sample project is an example of how to set up a double-observer study and specify analyses – see Sample Projects.

In this chapter we also provide some analysis guidelines, give a list of the output the engine can produce and cover various miscellaneous topics.

If you are familiar with the R software, you can run the DSM engine directly from within R, bypassing the Distance interface altogether. For more information, see Running the DSM Analysis Engine from Outside Distance.

There is an online help that accompanies the DSM R library. This contains more details about the methods and options available, and will be of use to users familiar with R. To open these help pages from within Distance, choose **Help | Online manals | DSM Engine R Help (html)**. The main function to examine is `dsm.fit()`.

Setting up a Project for DSM Analysis

The easiest way to set up a new project for an DSM analysis is using the Setup Project Wizard.

- In Step 1, under I want to: select Analyze a survey that has been completed.
- Be sure to tick the box indicating the Project will contain geographic information at the bottom of the Step 1 screen

- In Step 3, under Observer configuration, select Double observer. But see also Single Observer Configuration in the MRDS Engine.
- Follow through the rest of the wizard as usual.

Distance then creates the appropriate data fields for double observer data, and you can then import your data using the Import Data Wizard. For more about the data requirements, see Setting up your Data for MRDS Analysis.

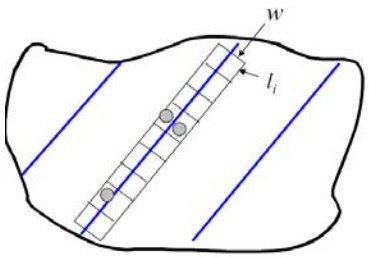
Alternatively, you can create the appropriate fields by hand, and manually create a new survey object with the appropriate observer configuration and data files. For more about survey objects, see Working with Surveys During Analysis in Chapter 7.

Setting up Your Data for DSM Analysis

It must be said that preparing your data is non-trivial. If you have not previously constructed data for a Distance project, you will be operating at a bit of a disadvantage. You will be well-served to closely study Chapter 5 of the Distance Users Guide (particularly the sections on Importing one file per data layer and Importing existing GIS data).

Segmenting your line transect data

There is no best way to accomplish this; some of the criteria for good segmentation will be dependent upon your dataset. Nevertheless, your transects will need to be divided into a number of "sub-transects" we will call segments. As a rule we can use to move forward, you will wish to have segments that are approximately equal in length to twice the truncation distance *w*, e.g., if you conducted a survey of ants and were able to detect them to a distance of 1 meter off the transect line, then you should divide your ant walking transects into segments roughly 2 meters in length. In this manner the resulting areas in which your survey has been divided is roughly square in shape.



Line transect after segmentation with individual segments of length l_i and truncation distance w.

Layer-specific data files

In conjunction with segmentation of your data, you must think about the hierarchal structure of your Distance project. The project we will be describing possesses this structure:

Study region
Transects
Segments
Observations

We will not be concerned with strata that may arise in some distance sampling applications. The sampling of our study region (otherwise know as the 'global layer' in Distance) consists of transects. Each transect is composed of segments

(thanks to the segmentation effort we performed previously). Within each segment we may have detections of the objects we are studying.

As you know from your previous work with Distance projects, each layer of a Distance project contains multiple fields. You will want to populate the Observation layer with data that you think may be influential in modelling the detectability of objects (e.g., observer, cluster size, etc.). This is exactly the type of modelling you have done with previous versions of the Distance software. You will also wish to populate the segment layer with data that you will wish to use in your spatial modelling (e.g., latitude, longitude, soil depth, prey biomass, etc.). Note these data are specific to the segment, so you will wish to think carefully about how to integrate data that is defined at a point so that it will be relevant at the spatial scale of a segment.

Constructing layer-specific files

If you surveyed your study area with sufficient effort, you have 20-40 transects (as suggested by Buckland et al. (2001)). Now that you have segmented each transect such that there are perhaps 10-20 segments per transect, the result will be 200-800 segments. It goes without saying it will be important to track the transect and segment within transect where each detection took place.

So, imagine a resulting dataset:.

Area	T-A	S-1	O-a
			O-b
		S-2	
		S-3	O-c
	T-B	S-4	O-d
		S-5	O-e
		S-6	
		S-7	O-f
			O-g
	T-C	S-8	
		S-9	O-h
			O-i
	T-D	S-10	O-j
		S-11	O-k
			O-l
		S-12	
	T-E	S-13	
		S-14	O-m
		S-15	O-n
			O-o

This represents 5 transects within the study area (labeled A through E), with Transect A comprised of 3 segments, Transect B comprised of 4 segments, C comprised of 2 segments, D having 3 segments, and E also having 3 segments). By coincidence there were also 15 observations (2 in segment 1, none in segment 2, 1 each in segments 3, 4 and 5, none in segment 6, 2 in segment 7, none in 8, 2 in 9, 1 in 10, 2 in 11, none in either 12 or 13, one in 14 and finally 2 in 15). Note, this depiction omits the covariates at either the segment or observation layer. If there are a large number of covariates included for consideration in modelling either the detection function or the response surface, the amount of data to bring into program Distance could be considerable.

We advocate the use of importing data by layers, which in this case might consist of 3 files (one for each layer "transect.txt", "segment.txt", and "observation.txt"). "transect.txt" would contain the transect label and its length, "segment.txt" would contain the label of the transect to which each segment belonged, along with segment-specific data such as segment length, latitude, and longitude mentioned earlier. Finally "observation.txt" would contain the identifier of the segment in which the detection took place, plus items such as the perpendicular distance of the detection, cluster size, and other data associated with detectability.

The observation layer file will need to contain three other fields beyond those already mentioned. If you are working with double platform (MRDS) designs, you will already include them in your data. If you are unfamiliar with double observer designs, and the letters MRDS mean nothing to you, then consult the section of this users guide regarding the MRDS engine. Setting up your Data for MRDS Analysis. There are two fields that will take the value '1' for each detection. They will be called 'Observer' and 'Detected' when they are imported. Finally the last field in this layer will be called 'Object' and it will contain a unique number for each object detected. However see also Single Observer Configuration in the MRDS Engine regarding elimination of these fields in favour of a simpler approach.

Producing a prediction grid in GIS

However, we are finished with the easy part of preparing our data for analysis with the spatial modelling capacity of Distance. The next step involves creation of an ESRI shapefile. You may have expected this step, because if you were going to perform spatial modelling it is only logical there would need to be a GIS in your future. So, take the plunge.

You will construct a prediction grid, because you will be predicting with covariates across your study region. Define some part of the universe that you call your study region (this requires considerable thought). Use your GIS system to produce a shape file of type *point*. These points should be thought of as the centers of cells into which you (courtesy of your GIS) have subdivided your study region.

Try **not** to construct a prediction grid with tens of thousands of cells; this will be agonizingly slow to import and analyze (when you are performing your final analysis with a dense prediction grid, be prepared to wait for considerable lengths of time for import, and particularly for bootstrap variance computations). Try not to exceed a couple thousand cells in your prediction grid.

For each of these cells you will need to populate an attribute table with all of the covariates you specified at the *segment layer* of your Distance project and that are included in the density surface model you wish to use for prediction. So, assuming you are using ArcGIS as your GIS software, perform the following:

- In ArcGIS, create a shape file of type point; containing covariates of interest; clipped to the extent of the study region,
- Compute the area of the cells (boundary cell sizes can be ignored for sufficiently dense prediction grid spacings relative to the size of the study region),
- Open attribute table of this object, and create a new field called “LinkID” of type "Number" and width 16.
- use the “Calculate Values” tool to fill that attribute field with a formula equivalent to the value of the FID field plus 1. This is

accessed by highlighting the newly-created field and pressing the right mouse button

  - Having made this modification the .dbf file you have worked with will also be modified (without explicit saving by you)

- Export the attribute table to ASCII. This file will be imported into the Distance project via the data import wizard in due course.

Importing prediction grid into Distance

This prediction grid exercise has resulted in the creation of a host of files. These will need to be mated with Distance to make the prediction grid available to the analytical tools in Distance. The steps involved in this process require considerable care. A refresher (or perhaps a first inspection) of the section of the Distance Users Guide Geographic (GIS) Data and Importing GIS Data by Copying an Existing Shapefile into the Data Folder would be helpful.

- A new layer must be created in your Distance project. Before creating a new layer, remember to ask Distance to either turn off the coordinate system for the global (study area layer), or change the default coordinate system in the Project settings (by selecting Data |Preferences from the menu).
  - You will need to have a shape file associated with the study area layer before creating the coverage layer in the following step. This is because the boundaries of the study region need to be known before Distance can try to drop a coverage layer atop the study area. If you wish to manufacture study area boundaries that represent a simple shape (such as a rectangle), this can be simply done by double clicking on the shape field (polygon) of the study area layer, and entering coordinates of four vertices.
- Create the new layer, making it a child of the Study Area layer (not a child of any other layers extant in the project). This layer will be of type ***coverage***.
  - This layer will need to have *exactly* the same number of records as the prediction grid you created with your GIS. If you have a small number of cells in the prediction grid, you can add records to the layer after you've specified spacing of cells so as to get approximately the correct number of records. With a large number of records in your prediction grid, you will wish to iterate the cell spacing such that you get close to the correct number of records, and then add or subtract records to arrive at the correct number of cells in your prediction grid.
  - It seems a bit of a shame, now that you've gone to all this work building this layer, but the layer created in this process will be overwritten by the GIS shapefile constructed earlier. It is critical that you note (write it down) the name of the layer containing the coverage grid you have constructed.
- Import the attribute data (exported to an ASCII file when you were working with ArcGIS). The attributes are only those fields containing covariates you may use to predict abundance throughout your study area. You will want to import data into the coverage layer of your Distance project.
- Go into the Data Folder for the Distance project you are constructing. Find the shape file (".shp"), and other files with the same prefix (but different extensions) for this coverage grid layer. Delete these.

- Copy the shapefile (and companion files) manufactured by your GIS work with the prediction grid into the Distance project data folder, renaming them identically with the name of the coverage grid data layer you have just manufactured.
- In this fashion, you have swapped the GIS-generated shapefile for the artificially manufactured coverage grid layer.
- Check to see that in the prediction grid layer, you have values in the shape (type point) fields: these are the coordinates of your prediction grid cells. Also make sure you have values in all rows for the attribute fields you have imported. As an extra check of the integrity of the prediction grid layer, you can ask Distance to map the prediction grid cells, by using the mapping function of Distance.

Defining DSM Models

Introduction to DSM Models

The concept underlying density surface modelling is that a response variable is modeled as a function of predictor variables. However, the response variable for our purposes can be a count, an estimated density, or an estimated abundance within each transect segment. For purposes of this users guide, we will focus upon estimated abundance as our response variable; with the estimated abundance having been adjusted by detection probability to account for individuals not detected during the survey.

To proceed, we will specify a response variable, a link function, and a distribution for error terms. Different link functions and error distributions are appropriate for different response variables. The remainder of this chapter will discuss the modelling of segment abundance, but for the sake of completeness we will depict some common combinations of response, link, and error distribution.

Response variable	Area of segment (offset)	Weighting	Link function	Possible error models
Abundance $\hat{n}_i$	Covered $a_i = 2wl_i$	None	Log	Quasi-poisson Negative binomial
Count n_i	Effective $\hat{a}_i = 2\hat{\mu}l_i$	None	Log	Quasi-poisson Negative binomial
Density $\hat{D}_i = \frac{\hat{n}_i}{a_i}$	None	Covered area $a_i = 2wl_i$	Identity	Normal Gamma

Specifying DSM Model Formulae

One must specify formulae that tell Distance which covariates to include in the model for fitting the density surface model to the predictor covariates. These covariates are attributes of the transect segment (or higher spatial scale). Aside: it does not make sense to include predictors at the observation level in models

that predict response at the segment level because there may be multiple observations within a given segment.

Formulae consist of a series of terms joined by operators such as "+". The terms represent covariates and the operators tell Distance how the covariates relate to one another.

For example, the formula

latitude + longitude + depth

means include the data from the fields latitude, longitude and depth as covariates.

The names of predictor covariates are possible transformations of the field names in the Distance project database. See the section Translating Distance Fields into DS and MR Covariates. The concepts of factor covariates (Factor and Non-factor Covariates in MRDS) also apply to the use of covariates for density surface modelling. Operators for formulae are the same as for MRDS modelling (); however there is the additional smoother operator s, that can operate either in a univariate (`s(depth)`) or bivariate (`s(latitude, longitude)`) manner. The smoother operator will fit a nonlinear spline. See Wood (2006) for details on fitting of GAMs to data.

To see some examples of model formulae, have a look at the Dolphins example project.

DSM Analysis Guidelines

These methods are relatively new, so we are only beginning to gain experience on effective analysis strategies. Some preliminary guidelines follow – please let us know if you can suggest some further guidelines or amendments.

Producing point and interval estimates of abundance in a study region using the DSM engine requires four steps

- Modelling the detection function,
- Modelling the estimated segment-specific abundance as a function of covariates,
- Extrapolation, courtesy of the model, from the covered region of the study region to the unsurveyed portion of the study region (given measures of the predictive covariates throughout the study region). This is termed the *prediction* step, and
- Producing variance estimates (and confidence limits) using parametric bootstrapping techniques.

We will not discuss the modelling of the detection function, as that is discussed elsewhere (Modelling the Detection Function). The remaining steps however, are unique to the density surface modelling, and receive attention herein.

Density surface model

The modelling of the response variable (abundance or count or density) proceeds by specification of the analysis that contains the estimated detection probability

for each object (individuals or cluster). The level at which the segments are represented in the Distance project is also required.

Specification of the form the relationship between the response and the covariates include the following:

- Generalized linear model or generalized additive model framework,
- Link function,
- Offset term (if necessary),
- Form of the error distribution, and
- Weighting factor for observations (if necessary);

along with the model formula that uses the operators described in Specifying DSM Model Formulae.

Modelling involves fitting a series of models to the data, and employing model selection to compare them. For GAMs, the model selection diagnostic metric is the generalized cross-validation (GCV) score. Models with small GCV scores tend to do best at predicting response when presented with a new set of covariates, and this is our objective when modelling density surfaces.

The number of knots in a GAM smooth (governing the wiggliness of the predicted relationship) is estimated from the data, with the default for a univariate smooth being 10 knots, and 30 knots for a bivariate smooth. The number of knots can be fixed by the user by stating `s(latitude, k=5, fx=TRUE)` as part of the model formula, where *k* is the number of knots.

Prediction of abundance to unsurveyed areas

Having modeled the relationship between the response variable and the predictor covariates within the covered region, we rely upon the model to predict the response in unsurveyed portions of the study region (hence the concept of these estimates being *model-based*). We use a prediction grid that partitions the study area into cells, with all relevant predictor covariates being measured within all cells. The estimated response in each of the cells can be aggregated (summed for abundance and count responses, averaged for density responses) to produce a study region estimate (or prediction to any other spatially-delineated region within the study region).

To perform this estimation, the necessary ingredients are

- Identification of the analysis that produced the preferred density surface model preferred for model-based estimates of the response variable.
- Name of the spatial layer in the Distance project that contains the prediction grid and the covariate predictors employed in the chosen density surface model for each cell (Importing prediction grid into Distance).
- Size of each cell in the prediction grid (this may be constant for all cells in the study area, or may be cell-specific). If the value is constant, the scalar value can be specified in the Distance form when the analysis is performed. Otherwise, a field in the prediction grid layer is populated with the size of each cell, and the name of that field is specified at the time of analysis (Producing a prediction grid in GIS).
- The regions over which the aggregation is to take place. Note you must specify a layer that contains polygons (because the edges of the polygon are necessary to determine prediction cells

inside/outside of the region). The polygons may be either part of the survey in the Distance project that gave rise to the sightings (such as strata), or some other polygon contained within the Distance project (an area of special interest such as a biological reserve, within the study region). The default is aggregation over only the study region.

Variance estimation using parametric bootstrap

Because density surface modelling consists of using estimated abundance or counts as the response variable, density surface modelling constitutes a form of meta-analysis (analyzing the results of analyses). As such, there are two components to uncertainty in the resulting estimates of abundance in the study region.

We can approximate the combination of these two forms of uncertainty by using the delta method approximation.

$$\{cv(\hat{N}_{overall})\}^2 = \{cv(\hat{p})\}^2 + \{cv(\hat{N}_{DSM})\}^2$$

The precision of the detection probabilities ($\hat{p}$) are provided by the engine that fitted the detection function. The uncertainty of the abundance estimate produced by the density surface model is determined by applying a parametric bootstrap using something called a moving block that shuffles the residuals from the fitted density surface model among segments.

This particular application of the parametric bootstrap is capable, in our experience, of producing rogue replicates that are orders of magnitude larger than the median replicate. This has the consequence of producing quite large estimates of variance for the density surface modelling component of estimation.

We have instituted a trimming algorithm attributed to Tukey (1977). Values that lie more than coefficient*interquartile range below the first quartile, or the same amount greater than the third quartile are considered outliers, and are not incorporated in the calculation of the quantile confidence interval for $\hat{N}_{DSM}$; likewise these outliers are not included in the computation of the $cv(\hat{N}_{DSM})$ used in the delta method approximation to produce the measure of precision that incorporates uncertainty both in the density surface model as well as the detection function model.

This is a new analysis engine – you can expect some teething problems. Contact the program authors if you can't resolve them (see Sending Suggestions and Reporting Problems).

Output from DSM Analyses

The DSM engine produces the following output:

- a summary of results in the **Analysis Browser**. For general information about the Analysis Browser, see the section Introduction to the Analysis Browser in Chapter 7. There are many results statistics available, and you can select which ones are shown independently for each analysis set using the Column Manager (see Column Manager Dialog in the Program Reference).
- a detailed listing of results in the **Results** tab of the **Analysis Details** window. These are described in the following section, MRDS Results Details Listing.

- a log of the analysis, highlighting any possible problems, in the **Log** tab of the **Analysis Details** window. For information about troubleshooting problems, see Chapter 12 - Troubleshooting.
- (optionally) plots from the results details can be imported into other programs.

DSM Results Details Listing

When an analysis has run, a great deal of information is available in the **Results** tab of the **Analysis Details** window. A separate result will be produced for the **fitting**, **prediction**, and **variance estimation** steps of a density surface analysis.

DSM fitting step

Results from this step are split into the following pages:

Response surface/summary

- A summary table showing the result of the fitting of the GAM (or perhaps GLM) showing estimated degrees of freedom for each term along with an overall percent of deviance explained by the model and generalized cross-validation score.

Response surface/Plot: gam-check

- Plots of various diagnostics (Q-Q goodness of fit plot, histogram of residuals, residuals against linear predictor, and response against fitted values)

Response surface/Plot: residual versus fitted

- A detailed (larger) plot of the residuals against the fitted values (rather than against the linear predictor so that the fitted values are on a scale comparable with the original per-segment estimated abundance). This is a way of examining residuals that we find most useful because it is depicted on a biologically relevant scale.

Response surface/plot: diagnostic plot

- As many of these plots are produced as there are smoothed predictor covariates in the model (non-smoothed predictors are not plotted). The response as a function of the predictor is shown, along with a confidence interval and rug plot showing values of the predictor variable for which there were observations.

DSM prediction step

Results from this step are split into the following pages:

Response surface/prediction: Estimation for entire study area

- Produces a single value, i.e., response aggregated over the entire study area

Response surface/prediction: aggregated sums

- Table of point estimates for each of the requested polygons within the study region for which an aggregated response was requested

DSM variance estimation step

Results from this step are split into the following pages:

Response surface/variance: Bootstrapped measure of precision

- Merely indicates version of the MRDS engine being used

Bootstrap variance computations: outlier removal

- Describes result of outlier removal (percent of replicates meeting the outlier criterion and list of those replicates values removed)

Response surface/Variance plot: bootstrap distribution

- Histogram showing distribution of bootstrap replicates after outlier removal

Bootstrap confidence interval for abundance within study area

- Percentile confidence limits incorporating only uncertainty associated with the density surface modelling,
- $cv(\hat{N}_{DSM})$ and point estimate of $\hat{N}_{DSM}$ and standard error derived from trimmed bootstrap distribution
- $cv(\hat{N}_{overall})$ from which a confidence interval around $\hat{N}_{overall}$ is calculated by back calculating a standard error from the CV, and applying a log-based interval around the point estimate

Bootstrap confidence interval for the sub-units specified

- Percentile confidence intervals from the bootstraps for each of the sub-regions requested.

DSM Analysis Browser Results

Currently, a relatively limited number of statistics are available in the Analysis Browser, for the density surface modelling step of the analysis. Statistics available for inspection via the Analysis Browser are: proportion of deviance explained, generalized cross-validation score, generalized cross-validation scale, and number of segments included in the analysis (including cells with response of zero). We will be expanding the list of outputs in future versions. There are no results available to the Analysis Browser for the prediction or variance estimation steps.

Exporting DSM Results

The methods of exporting results from the Analysis Browser and results pages of the Analysis Details to other programs are the same as those for CDS analyses, as documented in Exporting CDS Results in Chapter 8. For example you can copy the results details text by choosing **Analysis Results | Copy Results to Clipboard**, and you can copy plots by choosing **Analysis Results | Copy Plot to Clipboard**. Note though that in the case of the plot, the underlying data are not copied as well (unlike for CDS plots) – just the plot picture.

There is another way to get hold of the plots produced by DSM analyses: to directly access the image files produced when each analysis is run. Each plot that is displayed in the Results Details has a corresponding image in the project's R Folder. For more details, see Images Produced by R in Chapter 7 of the Users Guide.

Miscellaneous DSM Analysis Topics

Clusters of Objects in DSM

If the objects detected are clusters of individuals, you tell Distance this in the same way as for CDS analyses, in the Setup Project Wizard (see Survey Methods Wizard Page in the Program Reference).

Unlike CDS analysis, the Model Definition does not offer any regression methods for dealing with size bias. If you suspect size bias is a potential problem, the appropriate way to deal with it in an MRDS analysis is to include cluster size (or some transformation of cluster size) as a covariate in the detection function model(s).

The cluster size field is one of the fields with a fixed name in detection function formulae in DSM (see Translating Distance Fields into DS and MR Covariates) – in formulae you should use the name `size` regardless of the actual field name.

Stratification and Post-stratification in DSM

At the moment, the MRDS engine only accommodates one level of stratification, and this stratification is assumed to be geographic. There is no allowance for weighting responses among strata in any fashion.

Running the DSM Analysis Engine from Outside Distance

The DSM engine is implemented as a library in the free statistical software R. When you run a DSM analysis from Distance, Distance creates files containing a set of R commands and the appropriate input data, calls the R software, waits for the results and then reads them back in. For more about R, see R Statistical Software in Chapter 7 of the Users Guide.

Some users may wish to run the engine from outside the Distance interface – either from within the R GUI interface or from another program. For example, you may want to automate the running of analyses for simulations or bootstrapping.

To see the format of the input and data files produced by Distance, try running a DSM analysis in debug mode. To set debug mode on, choose **Tools | Preferences**, **Analysis** tab, and tick **Debug Mode**. When you run analyses in debug mode, the input and data files are created, but the analysis is not run. The Log tab displays the location of the files – these are created in a directory with a name "dst" followed by up to 4 numbers, located within the Windows temporary directory. The file "in.r" contains the commands. Data are located in the files "ddf.dat.r" (used in the detection function modelling), "region.dat.r", "sample.dat.r" and "obs.dat.r" (used in estimating density given a fitted detection function). You can use these files as templates for creating your own command and data files.

To run the analysis from within the R GUI (Graphical User Interface), you can cut and paste the commands from the file in.r. To run the analysis from another program, you can call R in batch mode – this is achieved by calling the program RCmd.exe, which is located within the /bin subdirectory of your R installation. For more details, see the R for Windows FAQ (in R, type `help.start()` and when a browser window opens, click on the FAQ for Windows port). For an example of its use, see the Log tab of any DSM analysis you have run that was not in debug mode – you should see a line of the form:

```
Starting engine with the following command:
C:\PROGRA~1\R\rw1091\bin\Rcmd.exe BATCH C:\temp\dst90474\in.r
C:\temp\dst90474\log.r
```

Users familiar with R may wish to work inside the R GUI. The DSM engine is contained in the library DSM. To load the library from within R GUI, type

```
library(dsm)
```

All the functions in the dsm library are documented – the main functions are `dsm.fit()` (fits the density surface model) and `dsm.predict()` (produces the estimated response across the prediction grid). You can open a copy of the help files from within Distance by choosing **Help | Online Manuals | DSM Engine R Help (html)**.

The use of these libraries in operating systems other than Windows is not supported (but may well work – let us know!)

Installing an Updated Version of the DSM Engine

The DSM engine is implemented as a library (called "DSM") in the statistical software R. From time to time, we may issue updated versions of the library, for example in response to reported problems. These will come as an archive file **dsm.zip**. To install the new version,

- Find the old version of dsm.zip in the Distance program directory.
- Copy the new dsm.zip over the top of it (you may want to rename the old version first, as a backup)
- Choose **Tools** | **Preferences** | **Analysis**, and tick the option to **Re-install analysis engine library on next** run.
- Open a project containing analyses that use the DSM engine (for example the dolphins sample project).
- Run one of these analyses.
- Check in the Log tab of the Analysis Details. You should find the line:
  `package 'dsm' successfully unpacked and MD5 sums checked`

The next topic describes how to check which version of the library is being used.

Checking Which Version of the DSM Engine is Being Used

The DSM engine is implemented as a library (called "DSM") in the statistical software R. From time to time, we may issue updated versions of the library, for example in response to reported problems. Therefore, before downloading a new version or reporting a problem you may want to check which version of the library is currently in use. To do this, re-run an analysis that uses the DSM Engine (such as one from the dolphins sample project) and look in the Log tab for the line

```
> library(dsm)
```

After it, you should see a line which looks something like the following:

```
This is dsm 1.0
Built: R 2.5.1; ; 2007-07-20 14:41:38; windows
```

If you are reporting a problem you should quote both the build number (in the above case 1.0) and the build date and time (2007-07-20 14:41:38).

The previous topic describes how to update to a newer version of the DSM Engine, if one is available.

When reporting results, you may want to cite the exact version (i.e., build number) of the library that used in the analysis. This is stored in the Log tab, as outlined above.

Fine-tuning a DSM Analysis

Advanced Topic

In most cases, the default options for estimation of detection function parameters in a DSM analysis work adequately. If you are a seasoned veteran in the use of `mgcv()` for fitting density surface models, you may wish to access some of the inner levers and knobs. This kind of fine-tuning is specified in the **Detection function | Control** page of the model definition.

To enter options on this page, type them in as a comma-delimited list. Any noninteger numbers should have a decimal point separating the integer and fractional parts – e.g., `38.98`. You can also use engineering notation – e.g., `3.898E1`. For some options (e.g., starting values), you need to specify a vector of numbers. To do this, write them out as a comma-delimited list prefixed by `c(` and suffixed by `)` – e.g., `c(4.7, 0.1, 0.2)`. Consult the R documentation for detailed instructions regarding the use of `gam.control()` features.

Aside! The options you specify are exported to R without change as arguments to `gam.control()`, which is why the above format is required.

Chapter 12 - Troubleshooting

Known Problems

A list of known problems at time of release is in the file ReadMe.rtf, in the Distance program directory.

For a more up-to-date list, see the Support page of the Distance web site (you can access the web site from Distance by choosing **Help** | **Distance on the web** | **Distance home page…**).

Internal Errors in the Interface

Internal errors occur when Distance encounters a situation it has not been programmed to expect. In response, Distance shows a message box that gives a technical description of the problem, as well as the place in the code that the problem occurred.

In some cases the cause of the problem is obvious. For example, if you use a database application to delete part of a distance project file and then open the file in Distance, internal errors will result. In most cases, however, an internal error indicates that you have found a program bug. If you suspect that the error is indeed caused by a bug, you should contact the program authors, detailing the exact circumstances under which the problem occurred and the exact text generated by the internal error message box. See Staying in Touch in Chapter 1 of the Users Guide for more information.

Recovering from Internal Errors

If you press the **OK** button on the internal error message box, Distance will try to continue as if the error had not occurred. In some cases it is successful, but in most cases further internal errors will be generated. If you are able, you should try to close the project file and exit Distance. If Distance will not allow you to exit, you should manually end the program using the Windows Task Manager. To do this, press the **Ctrl**, **Alt** and **Delete** keys at the same time, highlight the Distance program and press End Task. (You may have to press End Task again if a Wait message appears.) Now follow the instructions on the page Recovering from Unexpected Program Exit further on in this chapter.

Problems with the Analysis Engines

Errors and Warnings in the CDS and MCDS Analysis Engines

Some analyses may results in warnings or errors being generated by the CDS and MCDS analysis engine. The log tab will then be amber (warnings) or red (errors), and you should look in the analysis log to find the appropriate warning or error message. A complete list of these messages is given in the Appendix on the MCDS Engine Command Reference, under MCDS Engine Error and Warning Messages.

Internal Errors in the CDS and MCDS Analysis Engines

Occasionally, while running an analysis, the CDS and MCDS analysis engine encounters a situation that it was not programmed to deal with and either shuts down with an Internal Error message, or, more rarely, crashes with a Fortran run-time error. Neither of these will cause the Distance interface to crash – instead the analysis will be given error status, and any errors will be reported in the Log tab. If any results were generated they will be in the Results tab, but results stats are not presented in the Analysis Browser.

A list of internal errors is given in the Appendix on the MCDS Engine Command Reference, under MCDS Engine Error and Warning Messages. More about the output from a program crash is in that appendix under MCDS Engine Command Line Output. Note that output from a CDS program crash will only be saved if you ticked the option **Capture command line output from CDS and MCDS engines in WinNT** in the Analysis Preferences Tab of the Preferences Dialog.

Tip! Run-time errors are rather more common in the MCDS engine. If you are using Windows NT/2000/XP, then Distance saves a copy of the command-line output to the Log file. This can be useful in diagnosing the source of the problem, so you should make a note of what is recorded there when discussing the problem with the program authors.

If you are using Windows 98/ME, you can still access the command-line output, by running the analysis in debug mode within the interface, and then running the analysis using the analysis engine from the Windows command line. For more information, see Running the MCDS Engine in the Appendix on the MCDS Engine Command Reference.

Problems with the MRDS Engine

Please report any problems not listed below to the program authors.

Note! Before reporting any problems, please note down which version of the mrds library in R you are using, and the build date, and include this information with your problem report. To find this information, see the page on Checking Which Version of the MRDS Engine is Being Used in Chapter 10.

Problem	Solution
Plots cannot be viewed or are poor quality in Results tab of Analysis Details.	Change image format or image properties - see Images Produced by R.

Analysis did not produce any output, or ran with unexpected errors	Look in the Log tab at the R commands used and any error messages produced by R (see Analysis Details Log Tab). If there seems to be a problem with the data (e.g., a data selection query problem, it is worth running again with **Tools | Preferences | Analysis | Echo data to log** turned on. If you are an advanced user familiar with R, you could run the analysis again in Debug mode (see Running the MRDS Analysis Engine from Outside Distance), and then cut and paste the commands from the input file line by line. Once the problem has been encountered, you may be able to use your knowledge of R to determine the appropriate fix. If this is caused by a bug in the program, please contract the program authors.
Detection function estimates not produced due to problem in the fitting algorithm (e.g., lack of convergence)	You may need to manually specify starting values and bounds on parameters, or maximum iterations – see Fine-tuning an MRDS Analysis for details.

Stopping an Analysis

On occasion, you will want to stop an analysis that is running. For example, you may have set off a long bootstrap analysis by mistake. Or perhaps the analysis engine has locked up.

- To stop an analysis from the Analysis Browser, click on the **Reset Analysis** button .
- To stop the analysis from the Input tab of the Analysis Details window, click on the **Stop** button (which replaces the **Run** button when an analysis is running).

Warning! On some systems, although the analysis appears to have stopped, it actually continues running in the background. You can see this in Windows NT, 2000 and XP if you open the Windows Task Manager (from the Windows **Start** menu, choose **Run** and type in taskmgr, or click **Ctrl-Alt-Del** and choose the **Task Manager** option). If you look in the **Processes** tab of the task manager, for CDS and MCDS analyses you may see the process MCDS.exe still running after you clicked the **Stop** button in Distance. This is the CDS and MCDS analysis engine. To kill the process, highlight it in the Task Manager and click **End Process**. This may also happen when running the MRDS engine – in which case you will see Rterm.exe still running. You can end this process in a similar manner.

GIS Problems

If you are experiencing strange behaviour with a project that contains geographic data, it could be because the GIS data is invalid in some way. Symptoms include maps that are blank or for which **Full Extent** button in the Map Window doesn't seem to work properly, error messages when generating a grid layer or a design.

In these circumstances, the first option should be to check that the global and stratum data layers are valid (for examples of various possible geometry problems, see http://support.esri.com/index.cfm?fa=knowledgebase.techarticles.articleShow&d=26920).

For simple shapes, this is as simple as opening the Shape Properties Dialog for each polygon and checking the vertices are all there and in the appropriate order (for more on this, see GIS Data Format in Chapter 5). For more complex shapes (and for simple shapes as a double-check), it is worth running one of ESRI's shapefile checking tools. For example, there is a cleanshapefile macro available for Arc version 8 at http://arcobjectsonline.esri.com/default.asp?URL=/arcobjectsonline/samples/utilities/cleanshapefile/cleanshapefiles.htm, while for version 9 there is a "Check Geometry" tool available in ArcTools, under Data Management Tools | Features (see http://support.esri.com/index.cfm?fa=knowledgebase.techarticles.articleShow&d=27429).

A second possible problem is an inappropriate projection. For more on projections, see Coordinate Systems and Projections in Chapter 5.

Recovering from Unexpected Program Exit

Distance saves the changes you make to your project straight to the project files. This means that should Distance quit unexpectedly, for example due to a power cut or a fatal program error, all of your work is most likely to be safe. However, under some rare circumstances, your project can become corrupted. This may happen, for example, if Distance is part-way through making an update when the problem occurs.

Fixing a corrupted project

In Distance, the Project File and Data File are actually database files. If these files become corrupted, for example due to power failure while performing a write operation, it may be possible to fix them using the database repair functionality built into the compact database function. To do this, ensure all projects are closed in Distance, then select the menu item **Tools | Compact Project**. You will be prompted for the name of a project to compact. Select the damaged project and press **Compact**.

For more about compacting projects, see Compacting a Project in Chapter 4 of the Users Guide.

Appendix - Program Reference

Introduction to Program Reference

This part of the documentation is designed to be referred to whenever you want to know how to use a specific part of the Distance interface. Each section covers a different window in Distance.

For a more in-depth treatment of major concepts involved in using Distance, see the Users Guide.

Setup Project Wizard

The **Setup Project Wizard** guides you through the process of creating a new project. It is started by choosing **File | New Project…** from the main Distance menu.

At any stage during the setup process you can click on the **Next** button (or press **Alt**+**N**) to move on to the next step, or the **Back** button (**Alt**+**B**) to return to a previous step and amend previously selected options. Press **Cancel** (or **Alt**+**C**) to cancel the wizard and creation of the new project.

The first page prompts you for the type of project you want to create. You choice here will dictate the content and number of the next wizard pages. The choices are:

15. **Analyze a survey that has been completed.**
    Choose this option if you have performed a standard distance sampling survey, and want to use Distance to analyze your data. In the following pages the wizard will ask you for information about your survey. It will use this information to set up one survey object, and a simple data structure containing four data layers of type:Global, Stratum, Sample and Observation. If you want to set up a more complex data structure, you should choose option 5, below.
16. **Design a new survey.**
    Choose this option if you want to design a new survey, and Distance will create a global data layer, containing one record. You can then enter the co-ordinates of your study area using the Data Explorer. A more extensive design setup wizard is planned for a future version of Distance.
17. **Use an existing Distance project as a template.**
    Choose this option if you want to use an existing project as the basis for your new project. Distance will copy the project settings, data structure, survey objects, data filters and model definitions

from the project you select. The survey data and analysis results are not copied. You can then import your survey data and run the analyses. An example where this option is useful is where you have a set of standard analyses that you want to perform on several different datasets.

18. **Import a project or command file created in a previous version of Distance**
    Choose this option to import survey information and data from a Distance 2.2-3.0 command file, or Distance 3.5 project file, or to import all the information (including designs, surveys, analyses and results) from a Distance 4 project file.

19. **Exit the wizard and** set **up the project file manually.**
    Choose this option if you want to set up the project file by hand. Click **Finish**, to be taken straight to the **Project Browser**, from where you can create data layers and fields, survey objects, etc as required.

For more information about these options, see Creating a New Project in Chapter 4 of the Users Guide.

There is also a check box where you can specify whether the project will contain geographic (GIS) information. Your choice of option, above, dictates whether this check box is accessable – for example if you choose option 2 (design a new survey), then the project must be geographic.

After you have finished the wizard, and the project has been created, you can turn a non-geographic project into a geographic one by choosing **File | Project Properties** … and ticking the check box **"Project can contain geographic information"** on the Geographic tab.

Setup for Analyzing a Survey

This part of the Setup Project Wizard is for when you have performed a standard distance sampling survey, and want to use Distance to analyze your data.

In the following pages, Distance will ask you for information about your survey. It will use this information to set up one survey object, and a simple data structure containing four data layers of type Global, Stratum, Sample and Observation.

If you want to set up a more complex data structure, you should go back to Step 1 of the Setup Project Wizard and choose the option to set up the project manually.

Setup for Analyzing a Survey Introductory Wizard Page

This introductory screen gives some information about the part of the Set**up Project Wizard** for when you want to analyze a survey already completed.

To skip this screen in the future, tick **Don't show this introductory screen again**.

Survey Methods Wizard Page

In this screen the **Setup Project Wizard** guides you through the second step of project creation, which involves providing information about your survey methods.

The following information is required:

- The **type of survey**, which can be either a line transect, point transect or cue count survey.
- The **observer configuration**, which can be single or double observer configuration. Single observer is the norm for conventional distance sampling. Double observer configuration is used to estimate g(0) when it is expected to be < 1 – see Chapter 10 - Mark Recapture Distance Sampling for more on this.
- The **distance measurements** – that is the type of distances that were measured in the field. For line transects this can be perpendicular distances or radial distances together with the angle of the object relative to the trackline. For point transects and cue counts only radial distances are measured.
- The **type of observations** – this is whether recorded observations were of single individuals or clusters of individuals.

Note! NEW! For those used to Distance 4 and earlier, the sampling fraction option is now on the Multipliers Wizard Page..

Measurement Units Wizard Page

At this stage in the project setup you need to specify the measurement units for your data, from the drop-down list. These units are used when setting up the data structure.

Tip! You can change the measurement units later in the **Data Explorer** – double-click on the 4th header row of the field you want to change and select from the list of units.

Tip! You can measure data in one set of units, and output analysis results in another set. The units of measurement are specified here and in the data sheet. The units of analysis are specified in the **Units** tab of the **Data Filter**. See Data Filter Units in the Program Reference for more information.

Multipliers Wizard Page

At this stage in the project setup you need to specify the multipliers for your analyses. Multipliers are constants that are used to scale the density estimate. For more detail, see Multipliers in CDS Analysis in Chapter 8 of the Users Guide, and look up "Multipliers" in the index of the Distance book.

In this setup project wizard, you can add multipliers to account for the following situations:

- When the sampling fraction (proportion of the lines or points that were surveyed) is not 1. For example, imagine that only one side of the transect line was observed. In this case the sampling fraction will be 0.5. Another example is cue counting, where the sampling fraction is the proportion of a full circle that is covered by the observation sector. A third example is when all points or lines are visited multiple times – then the sampling fraction is the number of visits.

Tip! If the sampling fraction is not the same for all lines or points, you account for this by adjusting the survey effort at the data entry stage. For example, if you only sampled on one side of the line on 2 out of 20 transects, and all transects were 10km long, then you should enter a transect length of 5 km for those two transects. In this situation you set the overall sampling fraction to 1.

- Surveys where g(0) is less than 1.
- Cue count surveys (this box will automatically be checked if you selected cue counts in the Survey Methods screen).
- Indirect surveys - that is surveys where the animal of interest is not surveyed directly, but instead some object that is produced by the animal is surveyed. Examples include nest counts and dung counts. In this situation, two multipliers are added: one for the object production rate and another for the mean time to object disappearance. (Cue counts are a special case of indirect counts, where the decay time is instantaneous.)
- Other - this gives you the ability to define a generic multiplier at this stage.

Ticking the boxes associated with these options causes Distance to create a field for the multiplier in the Global data layer. You enter the multiplier value at the same time as your other data.

Most multiplier values are not known with certainty, but instead are estimates with associated standard error (SE) and degrees of freedom (DF). As well as creating a field for you to enter the multiplier value, Distance also creates fields to enter the SE and DF. An exception is the sampling fraction, which should be known. For the 'Other' multiplier, you can choose whether Distance should create the SE and DF fields or not.

Tip! You can also define multipliers after the project has been created, using the **Append Field** button in the **Data Explorer**. This gives you the ability to define as many multipliers as you like. However, if you add the multipliers in the Setup Project Wizard, Distance will automatically include them in the default Model Definition. These issues are discussed in more detail in the Users Guide page on Multipliers in CDS Analysis.

Finished Setup for Analyzing a Survey Wizard Page

This is the final stage of the setup project process prior to the creation of the new project database.

You are required to choose one of the following destination options:

- **Data Entry Wizard**: Go here if you want to enter your data from the keyboard and want to be guided through the process.
- **Data Import Wizard**: Go here if you want to import your data from a text file.
- **Return to Distance**: Choose this option if you want to enter your data using the **Data Explorer** (a non-guided interface that is similar to the **Data Entry Wizard** but is more flexible). You can still start up either the **Data Entry Wizard** or the **Data Import Wizard** by selecting them from the **Tools** menu.

Save your setup settings for the next time you setup a new project by ticking **Save current** set**tings as default**.

Setup for Designing Surveys

This part of the Setup Project Wizard is for when you want to design a new survey. Distance creates a global data layer, containing one record. You can then enter the co-ordinates of your study area using the Data Explorer. A more extensive design setup wizard is planned for a future version of Distance.

Setup for Designing Surveys Wizard Page

This window gives some information about the Setup Project Wizard option to setup a project for designing surveys. For more information about survey design, see Chapter 6 - Survey Design in Distance of the Users Guide.

Use Another Project as Template

This part of the Setup Project Wizard is for when you want to use an existing project as the basis for your new project.

Use Another Project as Template Wizard Page

Distance can use any Distance project as a template for the new project. Distance will copy the project settings, data structure, survey objects, data filters and model definitions from the project you select. The survey data and analysis results are not copied.

An example where this option is useful is where you have a set of standard analyses that you want to perform on several different datasets. For more information, see the Using an Existing Project as a Template in Chapter 4 of the Users Guide.

In this window, you select the project to import information from. You are then taken to the final page, where you choose whether to import the new data or type it in by hand.

Finished Use Another Project As Template Wizard Page

This is the final stage of the setup project process prior to the creation of the new project database.

You are required to choose one of the following destination options:

- **Data Import Wizard**: Go here if you want to import the new data from a text file.
- **Return to Distance**: Choose this option if you want to enter the data using the **Data Explorer** (a non-guided interface that is similar to the **Data Entry Wizard** but is more flexible). You can still start up either the **Data Entry Wizard** or the **Data Import Wizard** by selecting them from the **Tools** menu.

Save your setup settings for the next time you setup a new project by ticking **Save current** set**tings as default**.

Import from Previous Version of Distance

This part of the Setup Project Wizard is for when you want to import survey information and data from previous version of Distance.

Import from Previous Version of Distance Wizard Page

Distance can import options and data from the following previous versions:

- **Distance 2.0 to 3.0**.
  Under **Files of type**, specify “Distance 2.0 to 3.0 command files”, and

select the command file you wish to import.
Distance will use information from the **OPTIONS** section of the command file to create a Survey object containing information about the type of survey (line, point, etc), type of object (line or point), and type of distance measurements (radial or perpendicular). Distance will also create the appropriate data structure and import the survey data. Distance will not read the **ESTIMATE** section, so you will have to set up the analysis specifications again yourself.

If you are having difficulties importing a Distance 2.0 – 3.0 command file, the first thing to check is that it runs okay in the previous version of Distance. You can download Distance version 2.2 from the Distance web site if you do not have a copy.

- **Distance 3.5**.
  Under **Files of type**, specify "Distance 3.5 project files", and select the project file you wish to import.
  Distance imports the project settings, and uses them to create a Survey object. It will also create the appropriate data structure and import the survey data. Distance will not import the Data Filters, Model Definitions or Analyses – you will have to recreate these again manually.
- **Distance 4**
  Under **Files of type**, specify "Distance 4 project files", and select the project file you wish to import. Distance will import all of the information from the old project, including project settings, data, maps, designs, surveys and analyses.

Data Entry Wizard

The **Data Entry Wizard** guides you through the process of data entry. It has similar features to the **Data Explorer** (the **Data** tab of the **Project Browser**) but is more suitable for beginners as it guides you through the process of entering data, giving on-screen advice via a text box at the top of the window.

For new projects, you are taken to the **Data Entry Wizard** automatically from the end of the Setup Project Wizard if you choose the option **Proceed to Data Entry Wizard**. You can also access the **Data Entry Wizard** from the main Distance toolbar, by choosing **Tools** | **Data Entry Wizard**.

You can only access the **Data Entry Wizard** if you have a simple data structure, with a single global, stratum, sample and observation layer. In other cases, you must manipulate your data using the Data Explorer.

The introductory screen in the wizard provides you with an overview of the 4 data layers (Global, Stratum, Sample and Observation) in your project. If you want to find out more about the way that Distance stores survey data, you should read the Users Guide Chapter 5 - Data in Distance.

As with the other wizards in Distance, you navigate through the wizard by pressing the **Next** and **Back** buttons on the navigation bar (or pressing **Alt**-**N** and **Alt**-**B**).

If you do not want to see the **Data Entry Wizard** introductory screen again then tick the box **Don't show this introductory screen again**.

Global Layer Wizard Page

At the **Global** step of the **Data Entry Wizard** you should enter values for any multipliers you selected in the Set**up Project Wizard**, along with their standard errors. If you did not define any multipliers then you can move straight to the next screen.

The **Data Entry Wizard**'s screen is split into three parts. The upper section contains the help text to assist you at each stage of the data entry process. This section can be resized at any stage to make more space for your data. On the lower left there is a hierarchical view of the data layers in your project – this is the **Data Layers Viewer**. (The **Data Layers Viewer** is visible, but disabled in the **Data Entry Wizard** - but it's fully functional in the **Data Explorer**!) On the lower right there is a spreadsheet-like window, called the **Data Sheet**. Have a look at the **Data Explorer** help page to find out more about the **Data Layers Viewer** and entering data in the **Data Sheet**.

If you have not read Chapter 5 - Data in Distance in the Users Guide, you should probably do so before proceeding much further.

Stratum Layer Wizard Page

At the **Stratum** step of the **Data Entry Wizard** you should enter information about your survey strata. If your survey did not include stratification, you should enter the total survey area in the "Area" field, and then move to the next screen.

Note that if you do not know the area of your strata, or the total study area, you should enter 0. See Unknown Study Area Size in Chapter 8 of the Users Guide for more details.

If you have not read Chapter 5 - Data in Distance in the Users Guide, you should probably do so before proceeding any further. To find out more about stratification in Distance, see Stratification and Post-stratification in Chapter 8 of the Users Guide (although this is quite advanced!).

Have a look at the **Data Explorer** help page to find out more about the **Data Layers Viewer** and entering data in the **Data Sheet**.

Sample Layer Wizard Page

At the **Sample** step of the **Data Entry Wizard** you should enter information about your samples (transects or points).

If you want to find out about Data Field Types and other new Distance jargon, you should read Chapter 5 - Data in Distance of the Users Guide.

If you want to find out more about the **Data Layers Viewer** and entering data in the **Data Sheet**, have a look at the **Data Explorer** help pages.

Observation Layer Wizard Page

At the **Observation** step of the **Data Entry Wizard** you should enter information about your observations.

Have a look at the **Data Explorer** help page to find out more about the **Data Layers Viewer** and entering data in the **Data Sheet**.

If you have not read Chapter 5 - Data in Distance in the Users Guide, you should probably do so before proceeding any further.

To find out more about stratification in Distance, see Stratification and Post-stratification in Chapter 8 of the Users Guide (although this is quite advanced!).

Note! The Cluster Size column is of field type decimal, which means that non-integer cluster sizes can be entered. This could occur if, for example, cluster size is estimated by more than one observer and the mean is taken.

Finished Data Entry Wizard Page

To find out more about the data structure that you have created, see Chapter 5 - Data in Distance in the Users Guide.

To find out more about the **Data Explorer**, see the Program Reference page on the **Data Explorer**.

Import Data Wizard

This wizard guides you through the process of importing data into a Distance project file. The wizard can be started in one of two ways:

- from the last page of the Setup Project Wizard, by choosing the option **Proceed to Data Import Wizard**. This is the ideal way to import data into a new project.
- by selecting the menu item **Tools** | **Data Import Wizard**. This is the best way to add extra data from file into an existing project. You can also replace your existing data with the imported data - this is an option at the end of the Data Import Wizard.

General information about data import is given in the Users Guide pages on Data Import in Distance. Before you try importing data it is essential to have a good understanding of how Distance handles your survey data - such as would be gained by reading Chapter 5 - Data in Distance in the Users Guide.

The data import wizard has six screens. The first is introductory. The second asks you for the data source – the text file to import data from. The third allows you to specify the destination of the data – which data layers to put the data into, and how to assign rows in the text file to records in the Distance database. The fourth screen asks you to specify the delimiter used in the text file, and the fifth asks you to match up columns in the text file with fields in the database. The last screen allows you to check if the number of columns and rows is correct, and displays a log of any errors that occur during the import process.

Tip! After importing data, it is generally a good idea to carefully check it in the Data Explorer, even if the import seemed to go smoothly.

Warning! Importing large datasets into Distance takes a long time. We hope to improve the performance of the import routines in future releases. Meanwhile, if you have a very large dataset, consider importing it into Distance 3.5 (which was much quicker) and then importing the Distance 3.5 project into the latest version of Distance.

Data Source Wizard Page

At this step of the Import Data Wizard, you select the text file to import data from. For more information about the types of text file and required format, see the Users Guide pages on Data Import in Distance.

Note! Only files with the following extensions are allowed: txt, csv, tab, asc, htm, html. This is because of a security restriction in the database engine used by Distance. For more information, see the Microsoft knowledge base articles Q247861 and Q239105 (go to www.microsoft.com and click on support, then search).

Data Destination Wizard Page

At this step of the Import Data Wizard, you tell Distance where to store the imported data.

Destination data layers

Here, you specify which data layers the new data will be stored in. You choose the **Lowest data layer** and **Highest data layer**, and Distance shows you a list of the data layers where records will be added, and the parent of the highest data layer.

Normally, if you are importing data from a flat file containing all the survey data, the lowest data layer is an Observation layer, and the highest a Stratum layer. However there are other scenarios where you will want to make different selections. For example:

- when you are importing the data one layer at a time – see Importing one file per data layer in Chapter 5 of the Users Guide
- when there are no strata in your study, and you have already created a single stratum record
- when you have a complex data structure, with more than the default 4 data layers for analysis (Global, Stratum, Sample, Observation)

Location of new records

- **Add all new records under the first record in the parent data layer.** An example of when you would choose this option is when the highest data layers is a stratum layer (as the parent data layer is the global layer, which only contains one record).
- **Input file contains a column corresponding to the following field in the parent data layer.** Choose this option when you want rows from the input file to be assigned to specific records in the parent data layer. For example, imagine you are importing a text file into an observation data layer, containing one row for each sighting. Clearly, you want each observation to be assigned to the correct transect. Your data file contains a column for the sighting distance, and also a column with the transect ID. You choose this option, and under **Field name**, select ID. Another example of the use of this option is given in the Users Guide page Importing one file per data layer.

Creation of new records in lowest data layer

- **Create one new record for each line of the import file.** If your lowest data layer is an observation layer, this is usually the option you want to select. That way, if two successive records are the same (for example, two sightings at 0 distance in transect 1), then two records will be created for them in the data file.
- **Create new records only when the line differs from the previous line.** This option is useful when you have multiple rows that are the same in your input text file. For example,

imagine you are importing a file containing just the transect data, so that the lowest and highest data layer are both the sample layer. The semicolon-delimited text file has columns for stratum label, transect label and transect length:

```
Stratum A;Line 1A;10
Stratum A;Line 1A;10
Stratum A;Line 2A;10.3
Stratum A;Line 2A;10.3
Stratum B;Line 1B;5.7
Stratum B;Line 1B;5.7
Stratum B;Line 2B;8.4
Stratum B;Line 2B;8.4
```

there are 8 rows of data, but each transect is repeated twice. Choosing this option ensures that only 4 records are created in the sample data layer.

Data File Format Wizard Page

At this step of the Import Data Wizard, you specify the delimiter used to separate columns of the text file you are importing data from.

Options

Delimiter. The delimiter is the character used to separate columns of data. For more about the delimiters and data format, see Data Import in Chapter 5 of the Users Guide. Here, you choose the appropriate delimiter for your text file.

Ignore rows. Tick the box to allow Distance to ignore the first row of your text file. This is useful if, for example, you have put column names, or some other reminder, on the first row.

Sometimes, when coming to this page, you get a "Problem reading data file" message, and then no data is displayed in the bottom part of the wizard page. This is often caused by the wrong delimiter being selected by default – selecting the correct delimiter should fix the problem. If setting the delimiter doesn't solve the problem, then see Troubleshooting the Import Data Wizard.

If you commonly use this delimiter then on the Finished page, choose the option to **Save current** set**tings as default**, and your delimiter will become the default.

Data File Structure Wizard Page

This is the screen where you tell Distance which column in the import text file corresponds to which field in the Distance database.

There are two ways to achieve this: manually, by clicking on the first and second row of grey boxes (**Layer name:** and **Field name:**); or using one of the shortcuts.

Manually assigning columns to fields

You manually assign columns in the import text file to fields in the Distance database by clicking on the first and second row of grey boxes. In the first row you specify the Data Layer Name, and in the second, the Data Field Name.

For each column, click on the first row and choose from the drop-down list of data layer names. Press **Enter** to confirm the selection. Then click on the second row and choose from the list of available data field names. Once you

have assigned a field to one column it disappears from the list of available fields. (When you have assigned all of the available fields in that data layer to a column then further fields are created for you automatically, because Distance assumes you want to import additional fields of data - see Importing additional columns not yet in the Distance database, below.)

When you are using the drop down lists, pressing **Enter** confirms your selection and pressing **Esc** cancels the selection.

If you assign a field to a column by accident, simply click on the first row for that field, select "[Ignore]" and press **Enter**. This will clear the Data Field entry.

Ignoring a column in the import text file

By default Distance ignores all columns of text with the word "[Ignore]" in the data layer name row. By choosing Ignore, you can skip extra columns of text that you do not want to import.

Importing additional columns not yet in the Distance database

Imagine that you wish to import an additional column of data that identifies the sex of each animal observed into the observation layer. (You are going to use this field for post-stratification, so that you can see if detectability varies by species.) You have not already created a field for sex in the Data Explorer, so it is not on the list of available fields that drops down when you click on the second row. To specify a new field, you simply click on the row and type in a name for the field. In this case, because there are no remaining fields in the Observation layer, Distance automatically supplies the default field name "New Field"; however you may prefer to call the field "Sex". Distance automatically chooses Data Type "Text" (it will be stored as Text inside Distance) – choose from the drop-down list to assign another field type. It's important to get the field type right here as once the field is created, you won't be able to change it.

Using shortcuts to assign columns to fields

The following two checkboxes can be used to save time compared with manually assigning columns to fields in the Distance database.

Columns are in the same order as they will appear in the data sheet

If the columns in the text file correspond exactly to the order of the fields in the Data Explorer, then ticking this box automatically assigns the columns to the correct fields.

Before opening the Import Data Wizard, you can move fields in the Data Explorer by dragging and dropping them. You will also need to create any new fields before starting the Wizard.

First row contains layer names and field names of each column

In many database and spreadsheet packages, you can specify the contents of the first row when exporting data. To use this shortcut, you will need the first row to contain both the layer name and field name for each column, separated by some delimiter. For example, first row of the column corresponding to the field "Area" in the data layer "Region" would be "Region*Area", assuming that "*" is the delimiter used. Possible delimiters are: * | _ - and . (i.e., a full stop or period).

This option is most useful as part of streamlining data import from a database or other programmable application. Combined with the option to setup a new project using another project as template (see Using an Existing Project as

a Template in Chapter 4 of the Users GuideGuide,), it streamlines the application of a prespecified, standard set of analyses by relatively inexperienced users to a new set of data (such as might be required in a regulatory framework).

Finished Import Data Wizard

This is the last screen of the Import Data Wizard. Click **Finish** to import that data according to the options you have chosen. If there are problems during the import, a log of the errors, showing the row and column where they occurred, is displayed. If the problem cannot be readily corrected, see Troubleshooting the Import Data Wizard.

Options

Existing data: here you can choose to overwrite the existing data in the project, or append the new data to the current records.

Save current settings as default Tick this box to save the selections you have made in previous screens as the default for next time you run the Import Data Wizard.

Troubleshooting the Import Data Wizard

Only files with the following extensions are allowed: txt, csv, tab, asc, htm, html. This is because of a security restriction in the database engine used by Distance. For more information, see the Microsoft knowledge base articles Q247861 and Q239105 (go to www.microsoft.com and click on support, then search).

This page contains some hints that may help you if you're having problems importing data into Distance. For more general information about importing data in Distance, see Data Import in Chapter 5 of the Users Guide. For an overview of the Import Data Wizard, see theImport Data Wizard page of the Program Reference.

Problems between Step 3 (Data Destination) and Step 4 (Data File Format)

When you press the **Next** button, to move from Step 3 to 4, Distance tries to load the file you have selected and parse it using the default delimiter. Any problems that occur at this stage result in a "Problem reading data file" message being displayed.

Some of the possible messages generated by the database (last line of message box) are explained in more detail below:

- Error 3440: "An attempt was made to import or link an empty text file." This message typically indicates that the wrong delimiter is selected – selecting the correct delimiter should fix the problem. This message usually indicates that the file you chose to import was not a text file. You should check that it is a valid ASCII text file (for example by opening it in Notepad), and that it is correctly formatted for import (see Data Import in Chapter 5 of the Users Guide).
- Error 3047: "Record is too large." This message often indicates that Distance failed to recognize any end-of-line symbols in the file. Distance therefore assumes that the file is one line containing a large number of fields (columns) - too many to import. The Distance import engine expects the end of each line to be denoted by a Carriage return (Cr - ASCII 13) + Line feed (Lf - ASCII 10) combination. This is the standard used by almost all windows software packages. However, we have come across some cases in which certain versions of some packages can put either Cr+Cr+Lf

or just Lf at the end of each line. Lf alone is also the standard end-of-line symbol on Unix machines. If your file is not of the correct format, there are some possible remedies:

- Many software packages have different options for exporting text files - for example Text Only, MS DOS Text, Text with Line Breaks, etc. Try playing around with these options.
- Try opening the file in a different word-processor or text editing package and saving it as text. From our experience, we particularly recommend the shareware software TextPad (www.textpad.com), which has options to save files into PC format (**File format**, under **Save As…**), which automatically makes the line end in Cr+Lf.
- One user reported that the problem disappeared when they moved to a different machine, despite all the software settings apparently being identical.
- If the file comes from a Unix machine, FTP it to the windows machine as an ASCII file (not as a binary) - this will ensure that the Lf 's are replaced with CrLf (or use TextPad, above).

If none of these remedies work for you, or you are having ongoing problems, please contact the Distance Development Team.

- Error 3051: "The Microsoft Jet Engine cannot open the file ' '. It is already opened exclusively by another user, or you need permission to view its data." This message often occurs because the file is still open in the software package that you used to export it. Close the file in the other package and try again.

Problems after pressing Finish

If there are problems during the import operation, a box containing a list of error messages appears on the Finished page. These messages give the line number that the problem occurred on, and so are often helpful in pinpointing the cause of the problem.

Please let us know if you encounter a problem you cannot solve, or manage to troubleshoot a particularly tricky import problem. We can add your experience to the above list!

Project Properties Dialog

The project properties dialog box can be accessed from the main Distance menu by selecting **File** | **Project properties…**. It contains information about the current project. There are two tabs, General and Geographic.

General Project Properties Tab

The General tab of the Project Properties dialog displays information about the project file and associated data folder, including file sizes and locations. There is also a comments box in which you can enter some remarks about your project.

You can easily cut and paste from the comments box - right click on the box to see a pop-up list of options.

Geographic Project Properties Tab

The geographic tab of the Project Properties dialog allows you to view and change the geographic settings of the project. To find out more about geographic data in Distance, see Geographic (GIS) Data in Chapter 5 of the Users Guide.

Options

Project can contain geographic information.

If this box is checked, the project is a geographic one, and the data layers can be spatially referenced. Once a project is geographic, you cannot un-check the box. However, if a project is not geographic, you can convert it to a geographic project by checking the box.

Default coordinate systems

Use the options here to define the default coordinate systems for geographic data in the project, and for maps and geographic calculations. To find out more about coordinate systems, see Coordinate Systems and Projections in Chapter 5 of the Users Guide.

Geographic data. These settings will be applied to all new data layers created in the project. It is best to set these options before creating any data layers, as then all data layers will have the same coordinate system. Although this isn't strictly necessary (the only requirement is that they all have the same datum), it will make calculations quicker.

Maps and geographic calculations. These settings will be applied to all maps displayed in the project. In addition, they will be used as the default for survey designs. If the geographic data's geographic coordinate system is "[None]", or if the geographic data is already projected, then the projection is automatically set. In other cases, you can specify a map projection and projection parameters. Map units are the unit of distance used when displaying maps.

Conversion between coordinate systems

Densification is the process of adding vertices to lines when projecting them from one coordinate system to another. Densification tolerance is the maximum distance allowed before adding a new vertex. Distances are defined in terms of the geographic data units (usually latitude and longitude).

A densification tolerance of 0 (the default) means that no extra vertices will be added. Above 0, the higher the value, the longer the distance between new vertices and therefore the fewer the new vertices. Smaller values (above 0) mean more vertices are added – which takes more computer time but increases the accuracy of the projection.

The default densification tolerance can be set in the Preferences dialog – this default is applied to all projects. Because the optimal densification tolerance depends on the scale of the data and the accuracy required, it can be over-ridden here on a project-by-project basis if required.

Preferences Dialog

The Preferences dialog lets you set options that relate to the behaviour of Distance across all projects on this computer.

General Preferences Tab

Projects

- **Default project folder:** the default location used when creating and opening projects. Click **Browse**… to choose a new location.
- **Always create a backup copy when opening projects.** For more information about backups, see Saving and Backing up a Project in Chapter 4 of the Users Guide.
- **Check projects are on the local hard drive before opening them.** By default, when you open a project Distance checks that it is located on a local hard drive, as opposed to a removable drive, a network drive or a CD. If it is not on a local hard drive, it issues a warning. This is usually sensible, as performance is significantly degraded for projects not located locally. However, if you are sure you want to be able to access projects not stored on the local drive, then un-check this option.
- **Store results in compressed format in project databases.** By default, the information shown in the Results tab of the Details windows (e.g., Analysis Details) are compressed in the distance project database to make the project smaller. If this box is not checked, the results are stored as plain text. You will normally only want to un-check this if you are directed to do so by the program authors (for example, in response to a problem with the current compression routine that can occur in some versions of Windows such as Chinese, Japanese and Korean language versions).

Interface

- **Show Tip of the Day at startup.** If checked, the Tips window loads each time Distance is started. You can also open the Tips window by selecting **Help | Tip of the Day** from the main menu.
- **Save current position and size of all open windows as default for new projects.** Use this option to customize the look of new projects in Distance.
- **Default tab for runs with warning status.** In some types of runs, warnings are generated routinely – for example in MCDS runs with cluster size as a covariate. By default the Log tab opens when opening Analysis Details windows with Warning status, but you can use this option to bypass the Log tab and go straight to results.
- **Results and Log font size.** Use this option to set the default font size. You can also alter the font size by right-clicking in both the Results and Log windows and choosing the appropriate options.

Copy to clipboard

- **Column separator.** This option sets the separator used between columns when copying tables (e.g., Browser contents, data sheet, etc.) to the clipboard.
- **Row separator.** This option sets the separator used at the end of lines when copying tables to the clipboard. Most packages expect the end of line to be denoted by carriage return (Cr– ASCII 10) + line feed (Lf – ASCII 13). However, if you are experiencing trouble pasting tables into a package, try experimenting with other separators.

Geographic Preferences Tab

The geographic tab of the Preferences dialog allows you to view and change the default geographic settings. To find out more about geographic data in Distance, see Geographic (GIS) Data in Chapter 5 of the Users Guide.

Default coordinate systems for new projects

Use the options here to define the default coordinate systems for new geographic projects. To change the default coordinate system for a particular project, open the project and then go to the **Geographic** tab of the Project Properties dialog (**File** | **Project Properties…**).

To find out more about coordinate systems, see Coordinate Systems and Projections in Chapter 5 of the Users Guide.

- **Geographic data.** These settings are applied to new data layers.
- **Maps and geographic calculations.** These settings applied to all maps. In addition, they are used as the default for survey designs. If the geographic data's geographic coordinate system is "[None]", or if the geographic data is already projected, then the projection is automatically set. In other cases, you can specify a map projection and projection parameters. Map units are the units of distance used when displaying maps.

Conversion between coordinate systems

Densification is the process of adding vertices to lines when projecting them from one coordinate system to another. Densification tolerance is the maximum distance allowed before adding a new vertex. Distances are defined in terms of the geographic data units (usually latitude and longitude).

A densification tolerance of 0 (the default) means that no extra vertices will be added. Above 0, the higher the value, the longer the distance between new vertices and therefore the fewer the new vertices. Smaller values (above 0) mean more vertices are added – which takes more computer time but increases the accuracy of the projection.

The default densification tolerance is set here, but can be over-ridden on a project-by-project basis in the Project Properties dialog. This is because the optimal densification tolerance depends on the scale of the data and the accuracy required, which will vary between projects.

Maps

- **Scale factor for copying and exporting maps.** This setting determines whether the map image is enlarged or shrunk when copying a map to the clipboard. The default value is 1.0, which means the map is copied at the size it is displayed. Choose a value larger than 1 to enlarge the map and get a better quality image. Chose a value smaller than 1 to shrink the map and get a smaller image.

Survey Design Preferences Tab

Design engine

- **Echo commands to log.** When this box is checked, the commands issued to the design engine are written to the log tab of the design or survey details window. This is helpful for the developers for debugging purposes, but ordinarily should be unchecked to save project file space.
- **Time stamp log entries.** When this box is checked, each log file entry is accompanied by the time it occurred. This can be useful for recording how long different operations took.

Design details window - Coverage probability map.

- **Symbol size.** Allows you to change the size of the dots that make up the grid points. This is largely useful for producing nice looking maps to output.
- **Number of classes.** Determines the number of classes the coverage probability is split into.
- **Maximum and minimum values of classes.**
  - Setting the range of the classes to range from 0 – 1 can be useful when you are comparing between maps, as they are then all drawn on the same scale – however you will likely need lots of classes as most points will have similar coverage probabilities in many designs.
  - Setting the range of the classes between the minimum and maximum coverage for each design gives you the best chance to spot patterns in coverage probability for each design, but makes it harder to compare designs visually, as each will have a different scale.

Browser windows

- **Default columns for new browser** set**s.** Clicking the **Design Browser…** or **Survey Browser…** buttons opens the Column Manager dialog, and allows you to select default columns for new Design Browser or Survey browser sets.

Analysis Preferences Tab

Analysis Engines

- **Automatically lock data sheet whenever an analysis is run.** Normally, you don't want to change the data after starting with your analyses. Checking this box prevents you from inadvertently changing the data, because the data sheet is locked automatically after the first analysis is run. For more information, see Locking the Data Sheet, inChapter 7 of the Users Guide.
- **Echo commands to log.** When this box is checked, the commands issued to the design engine are written to the log tab of the design or survey details window. This is helpful for the developers for debugging purposes, but ordinarily should be unchecked to save project file space.
- **Time stamp log entries.** When this box is checked, each log file entry is accompanied by the time it occurred. This can be useful for recording how long different operations took.
- **Capture command line output from CDS and MCDS engines in WinNT.** When this box is checked, any output written to the command line when running MCDS.exe is captured and loaded into the log window. Generally, this only occurs if the engine crashes - saving this output can be useful for helping the program authors to diagnose the source of the crash. The option is selected by default, and only works in operating systems based on Windows NT (i.e. NT4, 2000, XP and subsequent OSs). For more about the output from program crashes, see MCDS Engine Command Line Output in the MCDS engine reference appendix.
- **Debug mode.** When this box is checked, running an analysis causes the temporary command files to be generated and placed in the Windows temp folder, but the analysis engine is not run. The

name and location of the command files is written to the Log tab, and the status set to Warnings (amber). This option is useful when debugging the analysis engine, or for making template command files.

R Software

For more details about the link between the R statistical software and Distance, see R Statistical Software in Chapter 7 of the Users Guide.

- **Folder containing R.** This contains the path to the R software. If you have R installed on your machine, this path is automatically read from the Windows registry the first time the Preferences dialog is opened, or the first time an MRDS analysis is run. You may want to update it when you install a new version of R – see Updating the Version of R that Distance Uses.
- **Properties of images generated by R**. Allows you to change various aspects of the image formatting and image file type. Clicking Image Properties… opens the R Image Properties Dialog.
- **Remove new objects that are created with each run**. When selected (the default) any new objects created during an analysis run are removed from the R objects file (called ".RData", this is stored in the R folder, below the project data folder). This keeps the R object file as compact as possible. However, under some circumstances you may want to keep the R objects – for example when they will be re-used in a subsequent analysis. For more details, see Using a Previously Fitted Detection Function to Estimate Density in MRDS in Chapter 10 of the Users Guide.
- **Re-install analysis engine library on next run**. When selected, running an analysis that uses R (either the MRDS or DSM engines) causes the appropriate R library to be re-installed from its archive in the Distance program directory (file mrds.zip or dsm.zip) into the current version of R before the analysis is started. This option is only selected if the library has become corrupted (unlikely), or if an updated version of the archive has been placed in the Distance program directory (e.g., the program authors have sent you a patched version of mrds.zip or dsm.zip). If selected, the option is automatically de-selected after the next run of R once the library has been re-installed.

Analysis Components window

- **Opening Data Filter and Model Definitions.** Both the Data Filter and Model Definition properties dialogs have lots of tabs. In many cases, you want to access the same tab repeatedly. To save time, you have the option here to go directly to the tab that was used last time you opened the dialog. Alternatively the dialog can open on the first tab, which was the default in previous versions of Distance.

Analysis Browser window

- **Default columns for new browser sets.** Clicking the **Analysis Browser…** button opens the Column Manager dialog, and allows you to select default columns for new Analysis Browser sets.

Project Browser

The Project Browser is the main interface to Distance projects. It opens automatically when a project is first opened, and is closed when the project is closed. Its' 6 tabs allow you to access the following parts of the interface:

- **Data**. The **Data Explorer** is the main interface to the data in your project. For more information about data, see Chapter 5 - Data in Distance in the Users Guide. For more about the Data Explorer, see Data Explorer in the Program Reference.
- **Maps**. The **Map Browser** allows you to create and manage maps of the geographic data in your project. If the project is not geographic, this tab is disabled. For more about geographic data, see Geographic (GIS) Data – Chapter 5 in the Users Guide. For more about the Map Browser, see Map Browser in the Program Reference.
- **Designs**. The **Design Browser** is for creating and managing survey designs. For more information about survey design, see Chapter 6 - Survey Design in Distance in the Users Guide. For more about the Design Browser, see Design Browser in the Program Reference.
- **Surveys**. The **Survey Browser** is for creating and managing survey objects. Surveys are generated by designs (see Chapter 6 - Survey Design in Distance in the Users Guide), and are also used as part of the analysis specification (see Analysis Components – Chapter 7 in the Users Guide). For more about the Survey Browser, see Survey Browser in the Program Reference.
- **Analyses**. The **Analysis Browser** is for creating and managing analyses. For more about setting up and running analyses, see Chapter 7 - Analysis in Distance in the Users Guide. For more about the Analysis Browser, see Analysis Browser in the Program Reference.
- **Simulations**. The **Simulation Browser** will allow access to the simulation capabilities of Distance – it is disabled at the moment but should be implemented in a future version of the software.

Data Explorer

The Data Explorer is the main interface for viewing and manipulating data in Distance. You access it via the **Data** tab of the Project Browser.

An alternative data interface is the Data Entry Wizard, which can be accessed from the Setup Project Wizard or by choosing **Tools** | **Data Entry Wizard**. The Data Entry Wizard has a more restricted interface, and can only be used with a simple data structure. For more information, see Data Entry Wizard in the Program Reference.

You cannot use the Data Explorer effectively until you understand the way that survey data is stored in Distance – make sure you've read the Chapter 5 - Data in Distance in the Users Guide before you continue!

The Data Explorer is split into three sections, the Toolbar, the Data Layers Viewer and the Data Sheet.

The Toolbar functions are summarized below. The other sections of the Explorer are discussed on the next pages.

The Toolbar

Buttons that Change the Way the Data Sheet Looks

- Compact View. Shows a compact view of the data. This makes the Data Sheet display only the ID and Label fields for the layers that are not currently selected. Compare the following picture with that for expanded view, below.

Study Area		Region		
ID	Label	ID	Label	Area
1	Stratify example	1	Ideal Habitat	85000
		2	Marginal Habitat	600000

- Expanded view. Shows an expanded view of the data, with all fields from all rows showing. In the example below we can see that there are actually g0 and g0 SE fields in the Global data layer, which were hidden in Compact view. Compact and Expanded view are mutually exclusive.

Study Area				Region		
ID	Label	G0	G0 SE	ID	Label	Area
1	Stratify example	0.8367	0.1738	1	Ideal Habitat	85000
				2	Marginal Habitat	600000

- Show Field Types. Shows or hides the Data Type and Modeling Type of each field. All of the examples above have shown the Data Type and Modeling Types hidden. The following example is with them shown. The first header row is the Data Layer Name, the second is the Field Name, the third is the Field Type, the fourth is the units and the fifth is the Source Database Type:

Study Area		Region		
ID	Label	ID	Label	Area
ID	Label	ID	Label	Decimal
n...	n/a	n...	n/a	nautmi2
Int	Int	Int	Int	Int
1	Stratify example	1	Ideal Habitat	85000
		2	Marginal Habitat	600000

The Data Sheet Lock/Unlock button

- Locks or unlocks the Data. When the data are locked, you cannot change or delete any cells. See Locking the Data Sheet in Chapter 7 of the Users Guide for more information.

Buttons for Managing Data Layers

- Create New Data Layer. Opens the Creates a new layer dialog, allowing you to create a new data layer.
- Delete Current Data Layer. Deletes the currently selected data layer.

Warning! If you delete a data layer, all data in the layer *and in all child layers* are lost. There is no undo button!

- Data Layer Properties. Opens the Data Layer Properties dialog, allowing you to view information about the layer.

Buttons for Adding and Deleting Fields.

- Insert Field Before Current. Inserts a new field before the current field.
- Append Field After Current. Appends a new field after the current field.
- Delete Current Field. Deletes the current field.

Warning! Proceed with caution, as deleting a field can prevent a Distance analysis from running properly!

Buttons for Editing the Data Sheet

For more details about the functions of these buttons, consult the pages on Editing, Adding and Deleting Records in the Program Reference.

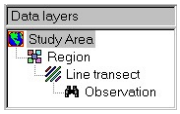
- Insert New Record Before Current. Inserts a new record before the current record.
- Append Record After Current. Appends a new record after the current record. Alternatively, you can press **Ctrl**+**Enter** or **Ctrl**-**Insert**.
- Delete Current Record. Deletes the current record. Alternatively, you can press **Ctrl**+**Delete**.

Warning! When you delete a row from a higher data layer then the corresponding rows from lower data layers also get deleted!

If you want to know more about the Data Explorer, proceed to next page on the Data Layers Viewer.

Data Layers Viewer

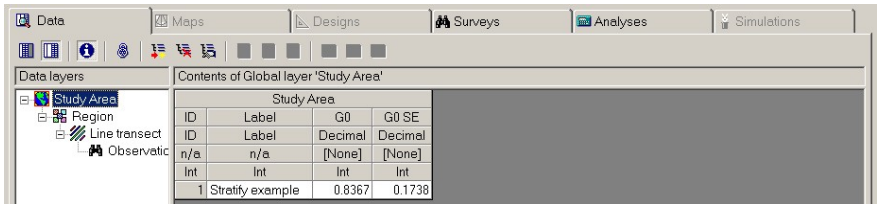
(For an overview of the data explorer, see the Program Reference page Data Explorer.)



Example of the Data Layer Viewer. In this case the project has 4 data layers.

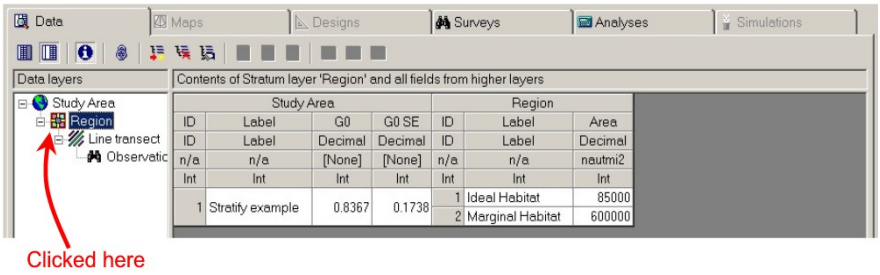
The Data Layer Viewer appears on the left in the Data Explorer. It presents a hierarchical view of the data layers in your project (see Chapter 5 - Data in Distance in the Users Guide for a discussion of the data layers). The icons by the data layer names indicate the Data Layer Type (see List of Data Layer Types in Chapter 5 of the Users Guide for a complete list).

Clicking on a data layer in the viewer shows the data for that layer in the Data Sheet, as well as data for all higher layers. When the Data Explorer first opens, only the Global Layer is displayed:



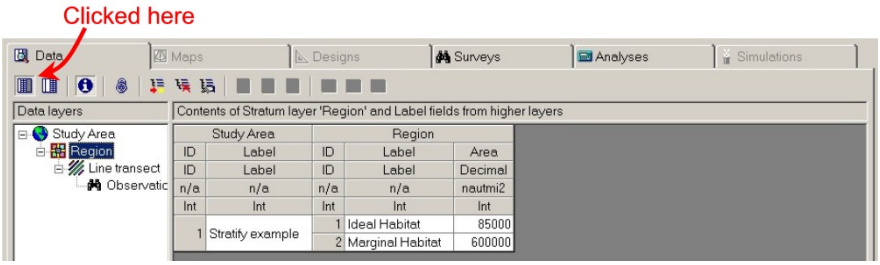
Part of the Data Explorer from the Stratify example project when first opened

If you click on the stratum data layer icon (in this case the stratum layer is called "Region"), the stratum data appears beside the global data:



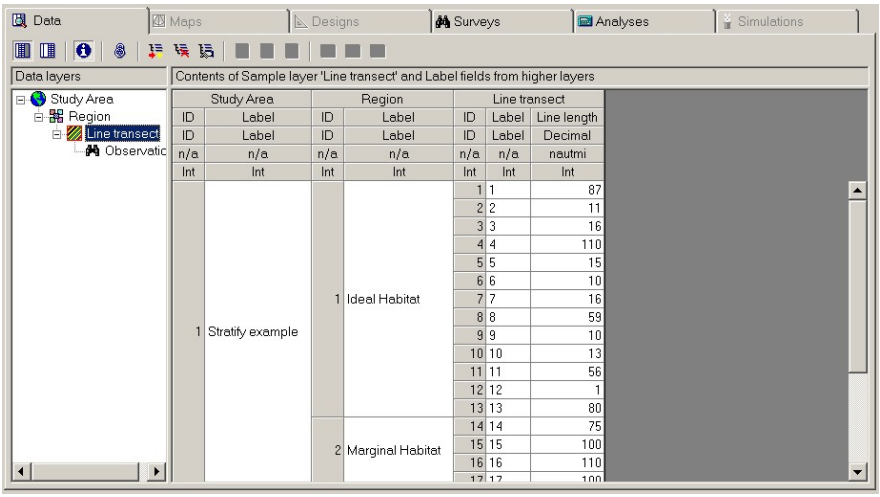
Part of the Data Explorer from the Stratify example project after the Stratum icon in the Data Layer Viewer has been clicked

You can compress the fields so that only the Label and ID column from the Global layer appears by clicking on the button on the toolbar (see Toolbar, above):



Same view as above, but with Compact View button clicked

Similarly, if you click on the sample data layer icon (in this case the sample data layer is called “Line transect”), the sample data appears beside the stratum data. Because the Compact View button is enabled, all fields in the Stratum data layer (Region) except the label and ID fields disappear from view:



Part of the Data Explorer from the Stratify example project after the Sample icon in the Data Layer Viewer has been clicked

Why go to all the trouble of selecting different data layers - why not just click on the observation layer and open up the whole data sheet at once? When all the data layers are open, the Data Sheet can become cluttered. It is often simpler, for example, if you want to see how many strata there are, to only open the strata data layer, and leave the sample and observation layer hidden.

Tip! When focus is on the Data Layer Viewer, you can use the arrow keys to move up and down the data layers, rather than clicking.

If you want to learn more about the Data Explorer, you should read the next page, which is about the Data Sheet.

Data Sheet

(For an overview of the data explorer, see the Program Reference page Data Explorer.)

The Data Sheet is intended to be as intuitive to use as any spreadsheet grid. However, the hierarchical nature of the data (observations within transects within strata within global) imposes some restrictions. In the following three pages, we describe how to perform the common tasks associated with manipulating survey data in the Data Sheet:

- Navigating Around the Data Sheet
- Editing, Adding, and Deleting Records
- Editing, Adding and Deleting Fields

Tip! You can easily copy your data from the data sheet to a spreadsheet or database package. Click on the Data Explorer to give it focus and then click on the **Copy to Clipboard** button on the main toolbar, or choose the **Data Explorer** menu item **Copy Data to Clipboard**. In your spreadsheet or database, choose **Paste**. You can use this Copy to Clipboard facility, together with the Data Import Wizard to provide a crude Import/Export facility for your data.

Sometimes you may run into problems pasting the data into your target package. This is usually caused by the symbol used by Distance to signify an end-of-row – you can change this in the **General** tab of the Preferences window.

Navigating the Data Sheet

(For an overview of the data explorer, see the Program Reference page Data Explorer.)

You can move around the Data Sheet by clicking on the grid and/or using the Up, Down, Left, Right Arrows as well as the Home, End, Page Down , Page Up and Tab cursor keys. Holding down Ctrl and a cursor key simultaneously, allows you to move in larger steps in that direction.

Tip! You can also move around by clicking on the grid and then moving the mouse while holding the left mouse button down. In this mode, if you move past the end of the visible grid, it scrolls for you.

Setting Data Sheet Column Widths

The data sheet columns widths are set automatically when the grid is first shown, based on the contents of the column. For all columns except ID, the width can be changed by clicking and dragging the grid lines in the column headers. Distance will then remember the column width you have set.

To find out more about the Data Explorer, consult the next page about Editing, Adding and Deleting Records.

Editing, Adding and Deleting Records

(For an overview of the data explorer, see the Program Reference page Data Explorer.)

Editing Records

The data grid has a distinct edit mode. The edit mode is entered by a mouse click or a key press. To replace the current contents of a cell, simply begin typing. To edit the contents, double click on the cell. The Left and Right Arrow keys are used within cells to move between individual characters and digits, and Home and End takes you to the beginning and end. Ctrl+Left and Right takes you to the beginning and end of the next word in the cell, while Shift+Left and Right selects part of the cell. Use Delete and Backspace keys as normal within the cell.

To exit edit mode, click on another part of the cell, use the Up or Down Arrow keys to get out, use the Left and Right keys to go past the contents of the cell, or use Page Up and Page Down.

Pressing Esc leaves edit mode without entering the data, returning the cell to its original state. Note you can also cut and paste into cells using the usual Windows keys, or right clicking the mouse button in edit mode. To delete the contents of a cell, press Delete (you will not be allowed to do this if the cell is required to contain a value).

You cannot edit the ID field, nor can you edit rows without an ID field (see Adding Records, below). Shape fields (in geographic projects only) are also special – if you double-click on a record in a shape field, the Shape Properties dialog appears.

Tip! You can tell whether you can edit a cell because the Focus rectangle (dashed box that shows you which cell is currently selected) turns from light dashes to heavy dashes when you can edit the cell.

Note! When you enter data into a field some data validation takes place to check that the data is of the correct type. This prevents you, for example, from entering text into a decimal or integer field, or entering decimal points into an integer field. Most cells must contain some value and cannot be left blank.

Tip! When you are in edit mode, the cell behaves just like any other text box. For example, you can right-click to bring up a pop-up menu containing useful commands such as Cut and Paste. You can also use the usual keyboard shortcuts (Ctrl-Insert, Shift-Insert and Shift-Del for Copy, Paste and Cut).

Tip! **Shift**+**Enter** takes you out of edit mode and on to the first field of the next record for that layer. Its designed to help with data entry - try it!

Adding Records

The Data Sheet is a representation of the underlying database, not a simple spreadsheet. Therefore, you cannot edit a cell in the Data Sheet unless there is a corresponding record in the database. For example, consider the following data sheet:

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1		
					3	Line 3	1		

The focus is on the observation cell that is below the record with ID 3 (you can tell this because there is a light focus rectangle on that cell). However, this cell has no ID in the ID field. This means that there is no record for this cell in the database. So, before you can enter the distance for this observation, you must add a record. You do this by clicking the Insert Record or Append Record buttons, or by typing the keyboard shortcut for Append Record, which is **Ctrl+Enter**. Insert Record puts a record before the current one, while Append Record puts one after the current one. In this case, since there is no record in the cell, they both have the same effect:

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1	4	0
					3	Line 3	1		

A record has been inserted with ID of 4. Notice that the focus rectangle is now heavy too, indicating that the value in the cell is editable. We enter the value 0.53 and hit **Enter**:

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1	4	0.53
					3	Line 3	1		

Now we wish to append another record, so we type **Ctrl+Enter** (the keyboard shortcut for Append record):

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1	4	0.53
								5	0.53
					3	Line 3	1		

Notice that the value of the record in the cell that previously had focus has been copied into the new record. We can now enter the next value for distance, 1.98 and hit **Enter**:

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1	4	0.53
								5	1.98
					3	Line 3	1		

We could continue this process until all the observations have been entered.

Note! If no objects were seen on a transect, then we do not add any records in the Observation data layer for that transect. For example, if there had been 3 objects seen on line 1, none on line 2 and 4 on line 3 then the completed data sheet would look like this:

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1		
					3	Line 3	1	4	5.43
								5	1.98
								6	0.89
								7	7.12

Adding Multiple Records at Once

Often it is more convenient to add more than one record at once. For example, in the above scenario we may know that there were 10 objects sighted on Line 2. We may wish to add all 10 records at once, and then type in the values for the distances. We can do this by double clicking on the ID field to bring up a

special window called the multi-record add dialog. This window allows you to add up to 99 rows at once.

Starting from the situation

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1	4	0
					3	Line 3	1		

we double click on the ID field, bringing up the multi-record add dialog:

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1	1	0
					3	Line 3	1		

In the text field, we type in 10: , and then click on either of the two buttons to the right of the up and down arrows (again, because there are no records associated with the cell, it doesn't matter which we press). Voilla!:

Study area		Region			Line transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Line length	ID	Perp distance
1	Mine site	1	No strata	100	1	Line 1	1	1	1.45
								2	9.54
								3	3.24
					2	Line 2	1	4	0
								5	0
								6	0
								7	0
								8	0
								9	0
								10	0
								11	0
								12	0
								13	0
								14	0
					3	Line 3	1		

We are now ready to enter the 10 values.

The multi-record add facility is particularly useful when you have interval (binned) data. Say you saw 50 objects in the first bin, which is between 0 and 10 meters. Insert a single record for the first object, and type in a distance of 5 meters (mid-way between the cutpoints). Then double-click on the ID field to bring up the multi-record add dialog. Enter 49 and press append. You one record gets copied into 49 new records, so now you have the 50 you want. See Interval (Binned/Grouped) Data in Chapter 8 of the Users Guide page on Interval Data for more details about how to enter and analyze interval data in distance.

This example used the Observation data layer, but of course it could have been done with any other layer.

Deleting Records

To delete a record, highlight the record in the Data Sheet and press the **Delete Current Record** button. You can only delete one record at once.

If you delete a record from the sample data layer, all of the observations associated with that sample will be deleted. Similarly, if you delete a record from the transect data layer, all of the samples and observations will be deleted. Remember, there is no undo button, so you may want to consider Backing up first!

To find out more about the Data Explorer, consult the last page in this chapter, Editing, Adding and Deleting Fields.

Editing, Adding and Deleting Fields

In many cases, you won't need to change the fields that Distance provides by default. However, there are a number of circumstances in which you may want to edit, add or delete fields – some of these are outlined in the relevant sections. Bear in mind if you are setting up the project to analyze data that it is usually best to make any changes to the fields before you start the analysis phase.

Editing Field Names

Distance provides default names for all the fields, but you may wish to change them. To edit a field name double-click on it in Data Sheet. You can also edit the Data Layer Names by double-clicking on them. If you edit the field names after you have created any Data Filters or Model Definitions, the results can be unpredictable!

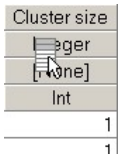
Editing Units

To change the units of a field, double click on the 4th row of the data sheet header for the field you want to change, and choose from the list of field types. n/a means "not applicable" – i.e., no units are required (e.g., the cluster size field doesn't require any units).

Moving Fields

You can change the order in which the fields are presented in the data sheet by using the mouse to drag the field name to a new location.

For example, the default order of fields in the observation layer for studies where objects are clusters is Perp Distance and then Cluster size. To swap the two fields, click on the Perp Distance column header with the left mouse button, hold the mouse down, move the mouse over the Cluster size column and release the mouse. As you drag the mouse you should see the cursor change from an arrow the drag pointer:



Adding Fields

There are many reasons for adding fields to the Distance database, beyond those provided by default. You may want to add extra multiplier fields in the global layer (see Multipliers in CDS Analysis in Chapter 8 of the Users Guide). You may want to add a field that will be used for post-stratification (see Chapter 8 of the Users Guide:Stratification and Post-stratification). You may want to add a column that defines a subset of the data that you will use to select data in a Data Filter, such as different species or years of data (see the Data Selection Tab of the Data Filter entry in the Program Reference). You may want to add fields for covariates in MCDS analyses, or you may be setting up a complicated data structure by hand.

To add a field to the data sheet, click on any cell in the appropriate data layer to give that layer focus. The click on either the Insert Field or Append Field button. Insert Field puts the new field where the current field is, and moves the current field to the right. Append Field places the new field after the current one.

After clicking the button, a small window will appear prompting you for the new field's name, Field Type and Units.

New fields are automatically filled with default values in the Data Sheet.

Deleting Fields

You may want to delete a field if, for example, you have defined a multiplier that you don't need any more, and is not being used by any Model Definitions. Another example is during survey design if you have deleted a design and want to remove the coverage probability field from the coverage layer. To delete a field, click on any cell in that field and click the Delete Field button. Distance will issue a standard warning, and will delete the field if you press OK.

Map Browser

The Map Browser allows you to create, sort, rename, delete and preview maps of the geographic data in your Distance project. It is accessed via the **Maps** tab of the Project Browser. The Map Browser is only accessible if the project is geographic. For more information about geographic data in Distance, see Geographic (GIS) Data in Chapter 5 of the Users Guide.

The Map Browser comprises a table showing a list of the maps that have been created, and optionally a preview pane, which gives a preview of the map that is currently selected. You can change the size of the preview pane by dragging the bar between the map table and preview pane.

Tip! For maps containing many shapes, the preview pane can take a while to draw. In these cases, it's better to hide the preview pane (click the button).

Note! Because of a limitation in the GIS engine Distance uses, the preview pane is blanked out (together with all other map windows) if any geographic data is changed.

To create a new map, click the button, or choose **Maps** | **New Map**. To rename the map, double click on the name in the map table, and type the new name. To view the map, so you can add layers, etc, select the map in the map table and click , or choose **Maps** | **View Map**.

Toolbar

- **Show Preview Pane**. Opens the preview pane on the right hand side of the Map Browser.
- **Hide Preview Pane**. Closes the preview pane.
- **Refresh Preview**. Causes the map in the preview pane to be refreshed, if it has become out of date (for example because a data layer has been deleted).
- **New Map**. Create a new map.
- **Delete Map**. Delete the currently selected map.
- **View Map**. Open the Map window for the currently selected map.
- **Up**. Move the currently selected map up in the map table.
- **Down**. Move the currently selected map down in the map table.

Design Browser

The Design Browser is the launching pad for survey design in Distance. From here you can create and run designs, arrange and sort them and view summaries of the results.

Before starting the survey design process, you should read Chapter 6 - Survey Design in Distance of the Users Guide.

Overview

The Design Browser is laid out like a spreadsheet, with one row for each design, and columns that give you useful information about the designs. The window is split into two panes. On the left there are columns of summary information about the design inputs: the status, unique ID number, design name, time created and time run. On the right are columns summarizing the results (which will be blank if the design has not been run yet) the actual columns showing can be customized using the Column Manager (press the button).

Tip! You can resize the panes by dragging the bar that divides them.

Tip! If you hold your mouse over a column header for a few moments, a small window pops up giving you an explanation of that column. This also works if you hold your mouse over a survey, data filter or model definition number: a window pops up giving you the name that corresponds with the number.

Designs can be grouped into Design Sets. A Design Set is a group of related designs – you are free to create, delete and rename sets, and choose which designs to group together. The current set name is listed after the word "Set:" on the design browser toolbar, and you can access a list of sets by clicking on the down arrow beside the current set name. You can create, delete and move sets using the buttons to the right of the current set name.

Tip! Transferring results to another application, such as a word processor is easy. Press the **Copy to Clipboard** button on the main toolbar or choose the **Designs** | **Copy Set to Clipboard**. This copies the contents of the current Design Set. In your word processor or spreadsheet, choose the Paste button.

Toolbar, and Designs menu

Set:

- **Set Name**. Gives the name of the current design set. Click on the name to edit it. Click on the drop-down arrow to get a list of other sets, from where you can click on another set to display its' contents.
- **New Set.** Creates a new Design Set.
- **Delete Set.** Deletes the current Design Set and all designs in it.
- **Arrange Sets**. Opens the Arrange Sets dialog, from where you can change the order that sets appear in the drop-down list of sets.

Design:

- **New Design**. Creates a new design.

Tip! The new design is based on the one that is currently selected in the Design Browser.

- **Delete Design**. Deletes the selected designs.
- **Design Details**. Opens Design Details windows for the selected designs
- **Run Design**. Runs the selected designs. For examples, see Example 3 - Automated Generation of New Surveys and Example 3 - Design Statistics in Chapter 3 of the Users Guide.
- **Reset Design**. Resets the selected designs. For designs that have been run, this deletes their Log and Results and returns the status to Grey (not run). For designs that are currently running, this cancels the run.

Tip! Sometimes it is useful to cancel an design while it is running - for example you may have set a very long coverage probability simulation running by mistake!

- **Move Design**. Moves the selected designs to another set. You are prompted for a set to move the designs to.
- **Arrange Columns**. This opens the Column Manager dialog for the current Design Set – see Column Manager Dialog in the Program Reference for more details.
- **Copy Set to Clipboard** (menu only). Copies the current set to the clipboard, from where it can be pasted into word processors, spreadsheets, etc.
- **Preferences** (menu only). Opens the Preferences dialog on the Survey Design page.

Tip! For the New Design, Delete Design and Design Details buttons, you can work with more than one design at once by highlighting multiple designs in the browser. To highlight more than one design, either:
(i) Hold the Ctrl key down and click on each design to highlight them.
(ii) Hold the Shift key and click on two non-adjacent designs to select all designs in between them.
(iii) Hold your mouse button down and move it over the designs you want to highlight, if they are adjacent.
(iv) Hold the Shift key and use the up or down keys to extend the current highlighting.

Tip! A shortcut way of opening the Design Details for a design is to double click on the status button in the right hand pane of the browser.

Sorting designs

To sort your design by any column, just click on the column header. One click sorts the column in ascending order, while another clicks sort in descending order. A little red arrow tells you which column is currently being used as the sort column, and whether it is an ascending or descending sort.

Survey Browser

The Survey Browser is the interface for managing survey objects in Distance. From here you can create and run surveys, arrange and sort them and view summaries of the results. Survey objects are used for two purposes: they are created from designs (as described in Chapter 6 - Survey Design in Distance in the Users Guide), and they are a component of an analysis (as described in Chapter 7 - Analysis in Distance of the Users Guide).

Overview

The Survey Browser is laid out like a spreadsheet, with one row for each survey, and columns that give you useful information about the surveys. The window is split into two panes. On the left there are columns of summary information about the survey inputs: the status, unique ID number, designID, survey name, time created and time run. On the right are columns summarizing the results (which will be blank if the survey has not been run yet) the actual columns showing can be customized using the Column Manager (press the button).

Surveys created from a design have a design number, and their status will be run (green), warnings (amber), or errors (red). Surveys used for analysis are usually not created from a design, and have no design number and status not run (grey). The right-hand results pane is therefore only useful for surveys created from a design.

Tip! You can resize the panes by dragging the bar that divides them.

Tip! If you hold your mouse over a column header for a few moments, a small window pops up giving you an explanation of that column. This also works if you hold your mouse over a survey, data filter or model definition number: a window pops up giving you the name that corresponds with the number.

Surveys can be grouped into Survey Sets. A Survey Set is a group of related surveys – you are free to create, delete and rename sets, and choose which surveys to group together. The current set name is listed after the word "Set:" on the survey browser toolbar, and you can access a list of sets by clicking on the down arrow beside the current set name. You can create, delete and move sets using the buttons to the right of the current set name.

Tip! Transferring results to another application, such as a word processor is easy. Press the **Copy to Clipboard** button on the main toolbar or choose the **Surveys** | **Copy Set to Clipboard**. This copies the contents of the current Survey Set. In your word processor or spreadsheet, choose the **Paste** button.

Toolbar, and Surveys menu

Set**:**

- **Set name**. Gives the name of the current survey set. Click on the name to edit it. Click on the drop-down arrow to get a list of other sets, from where you can click on another set to display its' contents.
- **New set**. Creates a new Survey Set.
- **Delete set.** Deletes the current Survey Set and all surveys in it.

- **Arrange sets.** Opens the Arrange Sets dialog, from where you can change the order that sets appear in the drop-down list of sets.

Survey:

- **New Survey**. Creates a new survey.

The new survey is based on the one that is currently selected in the Survey Browser.

- **Delete Survey**. Deletes the selected surveys.
- **Survey Details**. Opens Survey Details windows for the selected surveys
- **Run Survey**. Runs the selected surveys. Only surveys associated with a design can be run – see Chapter 6 - Survey Design in Distance in the Users Guide for more information.
- **Reset Survey**. Resets the selected surveys. For surveys that have been run, this deletes their Log and Results and returns the status to Grey (not run). For surveys that are currently running, this cancels the run.
- **Move Survey**. Moves the selected surveys to another set. You are prompted for the set to move the surveys to.
- **Arrange Columns**. This opens the Column Manager dialog for the current Survey Set – see Column Manager Dialog in the Program Reference for more details.
- **Copy Set to Clipboard** (menu only). Copies the current set to the clipboard, from where it can be pasted into word processors, spreadsheets, etc.
- **Preferences** (menu only). Opens the Preferences dialog on the Survey Design page.

For the New Survey, Delete Survey and Survey Details buttons, you can work with more than one survey at once by highlighting multiple surveys in the browser. To highlight more than one survey, either:
(i) Hold the Ctrl key down and click on each survey to highlight them.
(ii) Hold the Shift key and click on two non-adjacent surveys to select all surveys in between them.
(iii) Hold your mouse button down and move it over the surveys you want to highlight, if they are adjacent.
(iv) Hold the Shift key and use the up or down keys to extend the current highlighting.

A shortcut way of opening the Survey Details for a survey is to double click on the status button in the right hand pane of the browser.

Sorting surveys

To sort your survey by any column, just click on the column header. One click sorts the column in ascending order, while another clicks sort in descending order. A little red arrow tells you which column is currently being used as the sort column, and whether it is an ascending or descending sort.

Analysis Browser

The Analysis Browser is the launching pad for data analysis in Distance. From here you can create and run analyses, arrange and sort them and view summaries of the results.

Before starting your analyses, you should read Chapter 7 - Analysis in Distance in the Users Guide.

Overview

The Analysis Browser is laid out like a spreadsheet, with one row for each analysis, and columns that give you useful information about the analyses. The window is split into two panes. On the left there are columns of summary information about the analysis inputs: the status , unique ID number, survey number, data filter number, model definition number, analysis name, time created and time run. On the right are columns summarizing the results (which will be blank if the analysis has not been run yet) the actual columns showing can be customized using the Column Manager (press the button). Some additional notes on these columns are given under CDS Analysis Browser Results in Chapter 8 of the Users Guide.

Tip! You can resize the panes by dragging the bar that divides them.

Tip! If you hold your mouse over a column header for a few moments, a small window pops up giving you an explanation of that column. This also works if you hold your mouse over a survey, data filter or model definition number: a window pops up giving you the name that corresponds with the number.

Analyses can be grouped into Analysis Sets. An Analysis Set is a group of related analyses – you are free to create, delete and rename sets, and choose which analyses to group together. The current set name is listed after the word "Set:" on the analysis browser toolbar, and you can access a list of sets by clicking on the down arrow beside the current set name. You can create, delete and move sets using the buttons to the right of the current set name.

Tip! Transferring results to another application, such as a word processor is easy. Press the **Copy to Clipboard** button on the main toolbar or choose the **Analyses** | **Copy** Set **to Clipboard**. This copies the contents of the current Analysis Set. In your word processor or spreadsheet, choose the **Paste** button.

Toolbar, and Analyses menu

Set:

- **Set name**. Gives the name of the current analysis set. Click on the name to edit it. Click on the drop-down arrow to get a list of other sets, from where you can click on another set to display its' contents.
- **New set.** Creates a new Analysis Set.
- **Delete set.** Deletes the current Analysis Set and all analyses in it.
- **Arrange sets.** Opens the Arrange Sets dialog, from where you can change the order that sets appear in the drop-down list of sets.

Analysis:

- **New Analysis**. Creates a new analysis.

Tip! The new analysis is based on the one that is currently selected in the Analysis Browser (Half-normal / hermite in the picture above - when more than one are selected, look for the dashed focus rectangle around the one that has focus)

- **Delete Analysis**. Deletes the selected analyses.
- **Analysis Details**. Opens Analysis Details windows for the selected analyses
- **Run Analysis**. Runs the selected analyses. See the Running Analyses page in Chapter 7 of the Users guide for more information and tips.
- **Reset Analysis**. Resets the selected analyses. For analyses that have been run, this deletes their Log and Results and returns the status to Grey (not run). For analyses that are currently running, this cancels the run.

Tip! Sometimes it is useful to cancel an analysis while it is running - for example you may have set a large bootstrap analysis running by mistake!

Warning! On some systems, clicking on the Reset button while an analysis is running appears to stop the analysis, but it carries on running in the background, eating up system resources. For more about this, see Stopping an Analysis in Chapter 10 of the Users Guide.

- **Move Analysis**. Moves the selected analyses to another set. You are prompted for the set to move the analyses to.
- **Arrange Columns**. This opens the Column Manager dialog for the current Analysis Set – see Column Manager for more details.
- **Copy** Set **to Clipboard** (menu only). Copies the current set to the clipboard, from where it can be pasted into word processors, spreadsheets, etc.
- **Preferences** (menu only). Opens the Preferences dialog on the Analysis page.

Tip! For the New Analysis, Delete Analysis and Analysis Details buttons, you can work with more than one analysis at once by highlighting multiple analyses in the browser. To highlight more than one analysis, either:
(i) Hold the Ctrl key down and click on each analysis to highlight them.
(ii) Hold the Shift key and click on two non-adjacent analyses to select all analyses in between them.
(iii) Hold your mouse button down and move it over the analyses you want to highlight, if they are adjacent.
(iv) Hold the Shift key and use the up or down keys to extend the current highlighting.

Tip! A shortcut way of opening the Analysis Details for an analysis is to double click on the status button in the right hand pane of the browser.

Sorting analyses

To sort your analysis by any column, just click on the column header. One click sorts the column in ascending order, while another clicks sort in descending order. A little red arrow tells you which column is currently being used as the sort column, and whether it is an ascending or descending sort.

Note! Some columns, such as AIC and Delta AIC are special. When you click on these columns, your analyses are sorted by Data Filter first, and then AIC or Delta AIC column second. Why? AIC stands for Akaike's Information Criterion - an index of the relative fit of competing statistical models. The lower the AIC, the more parsimonious the model (other things being equal – look for AIC in the Distance Book). Delta AIC is the difference in AIC between the model with the lowest AIC and the current model. However, it only makes sense to compare AIC values, calculate Delta AIC and rank the analyses based on models fit to the same data. Sorting by the data filter column first ensures that this happens.

AIC is not the only model selection criterion that can be used. Distance also provides the following columns: AICc ("corrected" AIC), BIC (Bayes Information Criterion) and LogL (the log-likelihood). For more information about these criteria, and model selection in general, see Burnham and Anderson (2002).

Map Window

Map windows provide a view of the geographic data layers in a project. You can customize the map by choosing which data layers to include, and by panning and zooming around the map area. Any changes you make to a map are saved when you close the map.

You create maps in the Map Browser (see Program Reference, Map Browser for details). From there, click on the **View Map** button to open the Map window. You can have more than one map open at a time.

The map window is split into two panes. On the left is a pane containing map tools (currently just the layer control), and on the right is the map itself.

Note! Several features of the map window are not yet implemented. These include: the Info, Find and Spatial select map tools; the ability to customize the properties of each data layer, such as its colour, and add legends. We expect to implement these features in future releases.

Layer control

The layer control displays a list of the data layers currently shown on the map, with a legend showing the symbol used to display shapes on that layer. There is a tick box where you can turn off display of that layer. You can change the ordering of layers by clicking on a layer and dragging it above or below another layer.

Map tips

Map tips is a popup window that appears when you hover over a map feature, giving information about the feature. To enable map tips, click on the **Map Tips** button on the toolbar. A second row of tools opens on the toolbar, prompting you for the **Map Tip Layer** and **Field**. Select from the list of layers and fields, and then position the cursor over a feature on the map. The value of the selected field in that position will appear. For example, if you select a

sample layer “Line transect” and the “Label” field, then position the cursor over a particular transect, the map tip will display the label (name) of that transect.

The use of map tips can significantly slow down the map display, so turn them off when you’re finished!

Toolbar and Map menu

Tools in left-hand pane

- **Layer control**. Depressing this button displays a list of the layers in the map.
- **Info**. Not currently implemented.
- **Find**. Not currently implemented.
- **Spatial Select**. Not currently implemented.

About the map

- **Map Properties**. Opens the Map Properties dialog.
- **Refresh Map**. Refreshes the map – useful if some features have been edited so the map is out of date.

Manipulating layers

- **Add Layer**. Opens the Add Layer dialog, prompting you to choose the layer to add.
- **Remove Layer**. Removes the currently selected layer from the map.
- **Layer Properties**. Not currently implemented. This will allow you to customize the look of the data layer – for example the colour of the symbols.
- **Remove All Layers**. Removes all layers from the map.

Pan/zoom tools

- **Full Extent**. Zooms out to show the entire map area.
- **Zoom In**. When you click this button the cursor change to a “zoom in” symbol. Select on the map the area to zoom in to. To return the cursor to normal, click on the button again.
- **Zoom Out**. When you click this button the cursor change to a “zoom out” symbol. Click on the map to zoom out. To return the cursor to normal, click on the button again.
- **Pan**. When you click this button the cursor change to a “pan” symbol (a hand). Click on the map, and drag to move the display to another area. To return the cursor to normal, click on the button again.

Others

- **Map Tips** (toolbar only). select this button to turn on the display of map tips (see above for details).
- **Preferences…** (menu only). Opens the Preferences dialog at the Geographic tab.
- **Close** (menu only). Closes the map.

Design Details Window

This window is the main interface to each individual design. From here you can change the inputs, run the design to estimate probability of coverage or generate surveys, view the log file and the results pages. For more background about survey design, see Chapter 6 - Survey Design in Distance of the Users Guide.

To open a Design Details window, select the design in the Design Browser and click the **Design Details** button [picture here]. You can open up more than one design details window at once by selecting multiple designs before clicking the button. This could be useful if, for example, you want to view the results from two designs side-by-side. You can resize the design details window to fit your requirements by dragging the edges of the window (although there is a fixed minimum size). The last window you close sets the size for the next one to open.

A shortcut method of opening an Design Details window from the Design Details is to double click on an design's status button.

The Design Details window is divided into three tabs: Inputs, Log and Results. The tab that Distance first displays when you open an Design Details window depends on the status of the design. For designs that have not been run (grey status light in Design Browser), it opens the Inputs window. For designs that ran with warnings or errors (amber or red), it opens the Log tab. For designs than ran OK (green) it opens the Results.

Design Details Inputs Tab

Use the inputs tab to set up and run your design. It is divided into three sections: Design, Type of Design, and Comments.

Design

This section gives you some information about the design, such as the name, time created and time last run (i.e., time coverage probability estimated).

To change the name of the design, simply type a new name into the box labeled Name. Once you have typed the new name, hit Enter or click somewhere else on the window to apply the name to the design.

To run a design, click the **Run** button. This opens the Run Design dialog, where you are given the choice of either estimating coverage probability or creating a new survey. You cannot run a design until you have: (i) created a coverage probability grid in the project; (ii) selected the type of design; (iii) set the properties for the type of design. For more about the process, see Chapter 6 - Survey Design in Distance of the Users Guide.

Type of Design

To select the type of design, first choose the type of sampler (Line or point), and then the class of design. A list of all design types is given in Chapter 6 of the Users guide entitled Design Classes Available in Distance. A description and picture is given of the chosen class.

Before you run a design, you need to set the design properties, which you do by clicking on the **Properties…** button. This takes you to the Design Properties dialog.

Comments

The comments section is there for you to type some comments to yourself about the current design. For example, you might want to remind yourself of why you

chose to use these input parameters. The same section appears in the Results tab, so you can make comments about your results too.

You can give yourself more room by resizing the comments section. Put your mouse just above the Comments section header and dragging the section up and down. You may want to increase the height of the whole Design Details window (by dragging on its border) before you do this.

Design Details Log Tab

The **Log** tab is pretty-much identical for Design and Survey details, and both of them are covered here. This tab allows you to check any warnings or errors that occurred when you ran a design or generated a survey. Some messages in the log just report on the general progress of the design or survey run, thus if a warning or error does occur there is some indication of where in the sequence of the run this took place. Warning messages are coloured amber, and error messages red.

Warnings may indicate some problem with effort allocation, the definition of the stratum, or the sampler properties. Some warnings may also occur due to problems with the GIS component or due to problems with some geometric calculations for certain survey regions. This may possibly lead to problems calculating the design properties (which may then be invalid or very approximate) or generating a design (only part of a survey plan may be created). For example, a warning appears if no effort is allocated to any of the survey strata. If an incorrect stratum definition leads to the stratum having zero surface area, this is reported as a warning. If the sampler width or radius is large relative to the size of the survey region, then some coverage probability grid points will be covered by more than one sampler during a survey simulation. This may lead to coverage probability values greater than one. A warning letting you know that the coverage probabilities were constrained to fall within the [0,1] range will be displayed in this case.

Errors appear when an event occurs that completely thwarts the attempt to calculate design properties or generate an instance of a design. Such events occur when:

- An invalid description of the design is given.
- The project database cannot be accessed.
- The coverage probability grid layer cannot be located or the associated database table is missing.
- The coverage probability field name is invalid or cannot be validated.
- The temporary coverage probability grid layer cannot be created (the temporary layer is produced by projecting the original layer to the design coordinate system .
- The stratum layer cannot be located, it contains no strata, or the strata are incorrectly defined.
- The sample layer cannot be created or set up.
- The convex regions (required for the generation of some designs) data layer cannot be created or set up.
- The design axis coordinates used to orientate some designs are incorrectly defined.
- Problems with the GIS component or with some geometric calculations for certain survey regions lead to invalid designs.

- Too little effort has been allocated to a survey region to allow a particular design to be generated.

Design Details Results Tab

Header Information on the Design/Survey Results Tab

The **Results** tab for both designs and surveys displays some header information describing the design. Firstly, the design and sampler class are displayed. For surveys the name of the sampler layer is also shown. Some general properties to do with effort allocation are listed. The results show whether the coverage probabilities for a design run are simulation calculated or assumed even, and also which coverage grid is used, and the field where the coverage probability values are stored. For coverage probabilities estimated by simulation, the number of simulations is shown. A description of the stratum layers coordinate system type as well as that of the design is given. If the design coordinate system is projected, then the type of projection and its associated units are also given. The seed value used to initialize the random number generator (RNG) is also shown.

Coverage Probability Information on the Design Results Tab

If you selected the option that assumes the coverage probabilities to be even , approximate estimates of coverage probability are calculated analytically. For each stratum in the stratum layer the proportion of the sampled area relative to the survey region surface area is displayed. This proportion does not take sampler overlap into account, or that parts of the sampler may fall outside the study area along the region boundary.

If you opted to estimate the coverage probabilities by simulation, the design is created for the number of repetitions you specified. For each stratum in the stratum layer the minimum, maximum, mean, and standard deviation of the number of times each point in the coverage point grid is hit by the point or line sampler is shown. This is followed by the minimum, maximum, mean, and standard deviation of the coverage probability at each point in the grid. If the sampling intensity is very low the coverage probability values may be very small. If any of the coverage probability statistics are less than 0.001 then "< 0.001" is displayed instead of the value. It is always possible to retrieve the coverage probability values from the field in the coverage layer in which they are stored. You should aim for a design whose coverage probabilities are as even as possible, and the minimum number of hits for any design should be greater than zero.

To calculate coverage probabilities by simulation for designs using lines, each segment making up the sampler line is enclosed in a rectangle whose width is the same as that of the sampler, Similarly, for point designs, each point sampler is enclosed in a circle whose radius is the same as that associated with the point.

Design Property Information on the Design Results Tab

For each stratum in the survey layer the following general properties are displayed on the design **Results** tab:

- The number of point or line samplers that were specified on the effort allocation page and are expected to be generated.
- The samplers' radius or half-width.
- The expected sampler area coverage, which is the surface area covered by the samplers (generally less than will be realized in a survey).

- The surface area of the stratum and the expected proportion of the stratum covered by the samplers (based on the expected sampler area coverage).

Simple Random Point Sampling - Results Tab

The **Results** tab for both designs and surveys displays some header information for all survey designs.

Survey Plan Results

For each stratum in the survey layer the following are displayed:

- The number of point samplers that were specified on the effort allocation page and are expected to be generated.
- The actual number of point samplers generated and the associated sampler radius

The expected sampler area coverage, which is the surface area covered by the sampler points. Each point sampler line is enclosed in a circle whose radius is the same as that associated with the point. The area intersection of the circle is used in calculating the realized sampler area coverage. As the circles may fall partly outside the survey region the realized sampler area coverage is generally less the expected value. The potential overlap between the uniformly distributed point samplers is not taken into account when calculating the realized sampler area coverage.

The surface area of the stratum and the proportion of the stratum covered by the samplers.

Design Class Results

The design **Results** tab displays some general design properties and coverage probability information for all survey designs.

Systematic Point Grid Sampling - Results Tab

The **Results** tab for both designs and surveys displays some header information for all survey designs. For each stratum in the survey layer the following are displayed:

Survey Plan Results

For each stratum in the survey layer the following are displayed:

- The approximated number of point samplers displayed on the effort allocation page. This may differ from the actual number generated, as the points are generated according to the spacing specified for the systematic regular grid of sampler points.
- The actual number of point samplers generated and the associated sampler radius.
- The spacing between the systematic grid of points in the vertical and horizontal direction.
- The angle of the systematic point grid with respect to the x-axis, measured in an anti-clockwise direction from the positive x-axis.
- The expected sampler area coverage, which is the surface area covered by the sampler points. Each point sampler line is enclosed in a circle whose radius is the same as that associated with the point. The area intersection of the circle is used in calculating the realized sampler area coverage. As the circles may fall partly outside the survey region the realized sampler area coverage is generally less the expected value. The potential overlap between

the systematically distributed point samplers is not taken into account when calculating the realized sampler area coverage.

- The surface area of the stratum and the proportion of the stratum covered by the samplers.

Design Class Results

The design **Results** tab displays some general design properties and coverage probability information for all survey designs.

Parallel Random Line Sampling - Results Tab

The **Results** tab for both designs and surveys displays some header information for all survey designs. For each stratum in the survey layer the following are displayed:

Survey Plan Results

For each stratum in the survey layer the following are displayed:

- The number of line samplers that were specified on the effort allocation page and are expected to be generated. This number may differ from the number actually generated if the effort allocation is determined by line length rather than the number of lines. In this case the number displayed on the effort allocation page is an approximation.
- The actual number of line samplers generated and the associated sampler half-width.
- The total aggregated sampler length.
- The angle of the sampler lines with respect to the x-axis, measured in an anti-clockwise direction from the positive x-axis.
- The total length of the trackline, including distance spent off-effort moving from the end of one sampler to the beginning of the next one. Total cyclic trackline length includes the extra distance required to return from the end of the last sampler to the beginning of the first sampler.
- The expected sampler area coverage, which is the surface area covered by the sampler lines. Each sampler line is enclosed in a rectangle whose width is the same as that of the sampler. The area intersection of the rectangle is used in calculating the realized sampler area coverage. As the rectangles may fall partly outside the survey region the realized sampler area coverage is generally less the expected value. The potential overlap between the uniformly distributed line samplers is not taken into account when calculating the realized sampler area coverage.
- The surface area of the stratum and the proportion of the stratum covered by the samplers.

Design Class Results

The design **Results** tab displays some general design properties and coverage probability information for all survey designs.

Systematic Random Line Sampling - Results Tab

The **Results** tab for both designs and surveys displays some header information for all survey designs.

Survey Plan Results

For each stratum in the survey layer the following are displayed:

- The approximated number of line samplers displayed on the effort allocation page. This may differ from the actual number generated, as the lines are generated according to the spacing specified for the systematic sampler lines.
- The actual number of line samplers generated and the associated sampler half-width.
- The total estimated and realized aggregated sampler length.
- The spacing between the systematic line samplers.
- The angle of the sampler lines with respect to the x-axis, measured in an anti-clockwise direction from the positive x-axis.
- The total length of the trackline, including distance spent off-effort moving from the end of one sampler to the beginning of the next one. Total cyclic trackline length includes the extra distance required to return from the end of the last sampler to the beginning of the first sampler.
- The expected sampler area coverage, which is the surface area covered by the sampler lines. Each sampler line is enclosed in a rectangle whose width is the same as that of the sampler. The area intersection of the rectangle is used in calculating the realized sampler area coverage. As the rectangles may fall partly outside the survey region the realized sampler area coverage is generally less the expected value. The potential overlap between the uniformly distributed line samplers is not taken into account when calculating the realized sampler area coverage.
- The surface area of the stratum and the proportion of the stratum covered by the samplers.

Design Class Results

For each stratum in the stratum layer the minimum, maximum, mean, and standard deviation of the on-effort and total trackline length is displayed. The total trackline length is an approximation and does not assume that the observer will return to the start of the first sampler line, but rather that the survey is over after the last sampler line has been traversed. The distance covered to get to the first line sampler and back from the last line sampler is not considered. The design **Results** tab also displays some general design properties and coverage probability information for all survey designs.

Systematic Segmented Line Sampling - Results Tab

The **Results** tab for both designs and surveys displays some header information for all survey designs. For each stratum in the survey layer the following are displayed:

Survey Plan Results

For each stratum in the survey layer the following are displayed:

- The approximated number of line segment samplers displayed on the effort allocation page. This may differ from the actual number generated, as the line segments are generated according to the spacing specified for the systematic segment samplers.
- The actual number of line segment samplers generated and the associated sampler half-width.
- The total estimated and realized aggregated sampler length.
- The length of the line segments.
- The spacing between the systematic line segment samplers.

- The spacing between the tracklines along which the segments run.
- The angle of the tracklines with respect to the x-axis, measured in an anti-clockwise direction from the positive x-axis.
- The total length of the trackline, including distance spent off-effort moving from the end of one sampler to the beginning of the next one. Total cyclic trackline length includes the extra distance required to return from the end of the last sampler to the beginning of the first sampler.
- The expected sampler area coverage, which is the surface area covered by the sampler segments. Each sampler segment is enclosed in a rectangle whose width is the same as that of the sampler. The area intersection of the rectangle is used in calculating the realized sampler area coverage. As the rectangles may fall partly outside the survey region the realized sampler area coverage is generally less the expected value. The potential overlap between the uniformly distributed sampler segments is not taken into account when calculating the realized sampler area coverage.
- The surface area of the stratum and the proportion of the stratum covered by the samplers.

Design Class Results

The design **Results** tab displays some general design properties and coverage probability information for all survey designs.

Systematic Segmented Grid Line Sampling - Results Tab

The **Results** tab is the same as for the **Systematic Segmented Line Sampling** design, except that total trackline length is not calculated.

Equal Angle Zigzag - Results Tab

The **Results** tab for both designs and surveys displays some header information for all survey designs. For each stratum in the survey layer the following are displayed:

Survey Plan Results

For each stratum in the survey layer the following are displayed:

- The approximated line length displayed on the effort allocation page. This may differ from the actual length of the zigzag sampler generated, as the sampler is generated according to the angle specified for the equal angle zigzag.
- The actual length of the zigzag sampler.
- The number of zigzag segments generated (each determined by a change of zigzag direction) and the associated sampler half-width.
- The constant angle of the equal angle zigzag.
- The angle of the design axis, used to orientate the zigzag, with respect to the x-axis, measured in an anti-clockwise direction from the positive x-axis.
- The expected sampler area coverage, which is the surface area covered by the sampler lines. Each segment making up the zigzag sampler line is enclosed in a rectangle whose width is the same as that of the sampler. The area intersection of the rectangle is used in calculating the realized sampler area coverage. As the rectangles may fall partly outside the survey region the realized sampler area coverage is generally less the expected value. The potential overlap

between the rectangles is not taken into account when calculating the realized sampler area coverage.

- The surface area of the stratum and the proportion of the stratum covered by the samplers.

Design Class Results

The design **Results** tab displays some general design properties and coverage probability information for all survey designs.

Equal Spaced Zigzag - Results Tab

The **Results** tab for both designs and surveys displays some header information for all survey designs. For each stratum in the survey layer the following are displayed:

Survey Plan Results

For each stratum in the survey layer the following are displayed:

- The approximated line length displayed on the effort allocation page. This may differ from the actual length of the zigzag sampler generated, as the sampler is generated according to the spacing specified for the equal spaced zigzag.
- The actual length of the zigzag sampler.
- The number of zigzag segments generated (each determined by a change of zigzag direction) and the associated sampler half-width.
- The angle of the design axis, used to orientate the zigzag, with respect to the x-axis, measured in an anti-clockwise direction from the positive x-axis.
- The expected sampler area coverage, which is the surface area covered by the sampler lines. Each segment making up the zigzag sampler line is enclosed in a rectangle whose width is the same as that of the sampler. The area intersection of the rectangle is used in calculating the realized sampler area coverage. As the rectangles may fall partly outside the survey region the realized sampler area coverage is generally less the expected value. The potential overlap between the rectangles is not taken into account when calculating the realized sampler area coverage.
- The surface area of the stratum and the proportion of the stratum covered by the samplers.

Design Class Results

The design **Results** tab displays some general design properties and coverage probability information for all survey designs.

Adjusted Angle Zigzag - Results Tab

The **Results** tab for both designs and surveys displays some header information for all survey designs. For each stratum in the survey layer the following are displayed:

Survey Plan Results

For each stratum in the survey layer the following are displayed:

- The actual length of the zigzag sampler. The sampler is generated according to the length specified for the adjusted angle zigzag.
- The number of zigzag segments generated (each determined by a change of zigzag direction) and the associated sampler half-width.

- The angle of the design axis, used to orientate the zigzag, with respect to the x-axis, measured in an anti-clockwise direction from the positive x-axis.

The expected sampler area coverage, which is the surface area covered by the sampler lines. Each segment making up the zigzag sampler line is enclosed in a rectangle whose width is the same as that of the sampler. The area intersection of the rectangle is used in calculating the realized sampler area coverage. As the rectangles may fall partly outside the survey region the realized sampler area coverage is generally less the expected value. The potential overlap between the rectangles is not taken into account when calculating the realized sampler area coverage.

The surface area of the stratum and the proportion of the stratum covered by the samplers.

Design Class Results

The design **Results** tab displays some general design properties and coverage probability information for all survey designs.

Survey Details Window

Survey Details Inputs Tab

Survey objects are used for two purposes: they are created from designs (as described in Chapter 6 - Survey Design in Distance of the Users Guide), and they are a component of an analysis (as described in Chapter 7 - Analysis in Distance of the Users Guide). If you are using the survey as part of a survey design exercise, then you may want to Run the survey to generate an example sample data layer. If you are using the survey for data analysis, you will probably be most interested in setting or viewing the survey properties, via the **Properties…** button.

The Survey Details Inputs tab is divided into four sections: Survey, Survey Methods and Data, Design, and Comments.

Survey

This section gives you some information about the survey, such as the name, time created and time of the last run.

To change the name of the survey, simply type a new name into the box labeled Name. Once you have typed the new name, hit Enter or click somewhere else on the window to apply the name to the survey.

To run a survey, click the **Run** button. Note that you can only run a survey that is based on a design -- for more about the process, see Chapter 6 - Survey Design in Distance in the Users guide.

Survey Methods and Data

This section gives some outline information about the survey. To view more detailed information, and to edit the properties, click on the **Properties…** button. This will take you to the Survey Properties dialog.

Design

If your survey is based on a design, the design set and name will be displayed here. You can choose from the list if you want to change the design, although bear in mind that if the survey has been run (i.e., an example sample layer generated from the design), then the results will no longer apply, and the survey status will be reset to “not run”.

Comments

The comments section is there for you to type some comments to yourself about the survey. For example, you might want to remind yourself of why you chose to use these input parameters. The same section appears in the Results tab, so you can make comments about your results too.

You can give yourself more room by resizing the comments section. Put your mouse just above the Comments section header and dragging the section up and down. You may want to increase the height of the whole Survey Details window (by dragging on its border) before you do this.

Survey Details Log Tab

See the Log tab of the Design Details window for more information.

Survey Details Results Tab

See the Results tab of the Design details window for more information.

Analysis Details Window

Analysis Details Inputs Tab

You use the Analysis Details Inputs tab to set up your analysis. It is divided into five sections: Analysis, Survey, Data Filter, Model Definition and Comments.

Before you begin setting up and running analyses, you should read Chapter 7 - Analysis in Distance of the Users Guide, and the appropriate chapter after that relating to the analysis engine you want to use.

Analysis section of Analysis Details Inputs Tab

This section of the Analysis Details Inputs tab gives some information about your analysis, such as the name of the analysis, and the time it was created. You can also change the analysis name and run the analysis from here.

To change the name of the analysis, simply type a new name into the box labeled Name. Once you have typed the new name, hit Enter or click somewhere else on the window to apply the name to the analysis.

To run an analysis, click the Run button. See the Users Guide page in Chapter 7 on Running Analyses for more information and tips. Once the analysis has finished, it will automatically take you to the Results tab if it ran okay, or the Log tab if there were errors or warnings. While the analysis is running, the Run button changes to a Stop button. Press this to abort the analysis. (Pressing the Stop button has the same effect as pressing the Reset Analysis button for a running analysis in the Analysis Browser).

Survey section of Analysis Details Inputs Tab

In this section of the Analysis Details Inputs tab, you specify which Survey to use for the current analysis. For more about the use of surveys in data analysis, see Working with Surveys during Analysis in Chapter 7 of the Users Guide.

To find out more about the survey currently selected, you can click on the **Details…** button to open the Survey Details window for that survey.

Data filter section of Analysis Details Inputs Tab

In this section of the Analysis Details Inputs tab, you specify which Data Filter to use for the current analysis. For more background about Data Filters, see the Users Guide pages on Working with Data Filters and Model Definitions.

It is often easier to manipulate data filters (e.g., create new ones, delete them, rename them, etc. in the Analysis Components window – see the Analysis Components Window page in Chapter 7 of the Users Guide for more information).

The central window lists the Data Filters that are available for you to choose from. The one selected for the current analysis is highlighted on the list.

Choosing a different Data Filter for your analysis

If you want to choose another data filter for this analysis, click on the data filter you want. If you have results already for your analysis, Distance issues a warning that they will be deleted. This is because your results were generated with the old Data Filter and so will not correspond to your new choice. If you want to do a new analysis but keep your old results, then you need to go back to the Analysis Browser, click on the new analysis button and select the new data filter in Analysis Details window for the new analysis.

Making a new Data Filter

To make a new Data Filter press the **New…** button. A new filter will be created and appended to the current list. The new data filter is based on the data filter you have highlighted in the central window when you press the new button. The Data Filter Properties window is then opened up, so you can edit this new filter.

Scenario: Imagine you have run an analysis, and now want to try another analysis, but with just one part of the Data Filter changed (say a different truncation distance). Highlight the analysis you just ran in the Analysis Browser and click the **New Analysis** button. A new analysis is created, based on your current one. Now click the **Show Details** button to open an Analysis Details window. The old Data Filter will already be highlighted, so click the **New…** button to make a new Data Filter based on the old one. Make the changes in the Data Filter Properties and press **OK** to return to the Analysis Details window. Then click **Run** to run the analysis. Easy, eh!

Editing the Data Filter

Click the **Properties…** button. The Data Filter Properties window appears. Make any change you want in the Data Filter Properties and then press **OK** to return. Distance will warn you if you the data filter is associated with any analyses that have already been run.

Click **Properties…** if you just want to view the properties for this Data Filter, rather than edit them, and then press **Cancel** in the Data Filter Properties window to return without saving any changes.

Double clicking on the ID of a Data Filter in the central window is a shortcut way of opening the Data Filter Properties for that filter.

Renaming a Data Filter

Click on the data filter name, and start typing.

Deleting a Data Filter

To do this, go to the Analysis Components window.

Model definition section of Analysis Details Inputs Tab

In this section of the Analysis Details Inputs tab, you specify which Model Definition to use for the current analysis. For more background about Model Definitions, see - Working with Data Filters and Model Definitions in Chapter 7 of the Users Guide.

It is often easier to manipulate Model Definitions (e.g., create new ones, delete them, rename them, etc. in the Analysis Components window – see the Analysis Components Window in Chapter 7 of the Users Guide for more information).

The central window lists the Model Definitions that are available for you to choose from. The one selected for the current analysis is highlighted on the list.

Choosing a different Model Definition for your analysis

If you want to choose another Model Definition for this analysis, click on the Model Definition you want. If you have results already for your analysis, Distance issues a warning that they will be deleted. This is because your results were generated with the old Model Definition and so will not correspond to your new choice. If you want to do a new analysis but keep your old results, then you need to go back to the Analysis Browser, click on the new analysis button and select the new Model Definition in Analysis Details window for the new analysis.

Making a new Model Definition

To make a new Model Definition press the **New…** button. A new filter will be created and appended to the current list. The new Model Definition is based on the Model Definition you have highlighted in the central window when you press the new button. The Model Definition Properties window is then opened up, so you can edit this new filter.

Editing the Model Definition

Click the **Properties…** button. The Model Definition Properties window appears. Make any change you want in the Model Definition Properties and then press **OK** to return. Distance will warn you if you the Model Definition is associated with any analyses that have already been run.

Click **Properties…** if you just want to view the properties for this Model Definition, rather than edit them, and then press **Cancel** in the Model Definition Properties window to return without saving any changes.

Double clicking on the ID of a Model Definition in the central window is a shortcut way of opening the Model Definition Properties for that filter.

Renaming a Model Definition

Click on the Model Definition name, and start typing.

Deleting a Model Definition

To do this, go to the Analysis Components window.

Comments section of Analysis Details Inputs Tab

The comments section is there for you to type some comments to yourself about the current analysis. For example, you might want to remind yourself of why you chose to use these input parameters. The same section appears in the Results tab, so you can make comments about your results too.

You can give yourself more room by resizing the comments section. Put your mouse just above the Comments section header and dragging the section up and down. You may want to increase the height of the whole Analysis Details window (by dragging on its border) before you do this.

Analysis Details Log Tab

This tab of the Analysis Details window allows you to check any warnings or errors that occurred when you ran an analysis.

If you run an analysis and there are warnings, the Log tab is colored amber. If there are errors, this tab is colored red. In either case, when you open the Analysis Details window, this tab will display on top by default.

You can change the default tab for analyses that ran with warnings to the Results tab. The option is under the **Analysis** tab of the Preferences window. This option is useful if you are running analyses that regularly generate warnings, but where you want to disregard the warning message.

The Log tab is split into two sections. The top contains the analysis log – a list of the commands that Distance used to do your analysis, together with the message that Distance sent back while executing the commands. The bottom section gives a summary of any warning and error messages. You can resize the bottom section by clicking just above it and dragging.

To quickly go to the part of the analysis Log that contains a warning or error message, click on that message in the bottom section of the Log tab.

To save time, Distance only indexes and colors the first 30 warnings and 30 errors.

One common reason for analyses returning an error status is that the data selection criteria in the Data Filter have been entered incorrectly. To check what data are being sent to an analysis, tick the "Echo data to log" option in the Analysis tab of the Preferences dialog.

To copy the log file to the clipboard, click the **Copy to Clipboard** button on the main toolbar, or choose the **Analysis - Log** menu option **Copy Log to Clipboard**.

You can change the font size of the Log text in the General tab of the Preference dialog.

Analysis Details Log Tab - CDS and MCDS

Users of Distance 3.0 and earlier will recognize this as the old Log file. To help you find the important messages, warnings are colored Amber and errors Red.

Unfortunately, many of the messages are rather cryptic, and people who did not use Distance 3.0 and earlier will not be able to understand the command syntax. The good news is that we expect warnings and errors to occur rather rarely, at least for conventional distance sampling analyses, because the graphical user interface in the new versions of Distance do not allow you to make many of the mistakes that Distance 3.0 and earlier allowed.

As a check, you may want to see a log of the data that was output to the Distance analysis engine. To have the data echoed in all future analyses, tick the check-box in **Analysis** of the Preferences dialog (under the menu item **Tools** | **Preferences**).

Errors and warnings during bootstrapping

Distance treats errors and warnings that are generated during bootstrap estimation of variance differently from normal errors. These errors and warnings are labeled as "Bootstrap Error:" and "Bootstrap Warning:", and they are all colored amber. If an error occurs during the bootstrap, the status is not set to error (red), but to warning (amber) - this is because the error only affects the variance estimation, not the whole analysis.

Internal errors

On occasion, the analysis engine's internal algorithms may fail. In this case it usually generates an "Internal Error" in the Log, and the analysis status is set to error (red). In extreme cases, the analysis engine may crash. For more information about this, see the Troubleshooting – Chapter 12 of the Users Guide, particularly the page on Internal Errors in the CDS and MCDS Analysis Engines.

Analysis Details Log Tab - MRDS

As for CDS and MCDS analyses, the analysis log starts with a list of the data selection queries.

It next details how it has translated the fields names in the Distance project database into names that can be used in detection function formulae. This is a very useful part of the log to refer to when building new model formulae, as incorrect covariate names are a common source of problems in the analysis. For more on this, see Translating Distance Fields into DS and MR Covariates in Chapter 10 of the Users Guide.

The log then gives the command used to start the R statistical software in batch mode, followed by the R commands used and any response from the R software. This section is very useful for diagnosing any problems, although the error messages from R are sometimes quite crypic.

The log finishes by reporting the run status returned by R.

Analysis Details Log Tab – DSM

Analysis Details Results Tab

This tab on the Analysis Details Window allows you to view the results of your analysis in detail.

The Results tab is divided into two sections: Results pages, and Comments.

At the top of the results pages, there is a drop-down list of all the pages that are available. You can navigate through the pages by choosing from this list or using the **< Back** and **Next >** buttons. Detailed information about the contents of the results pages is given under Output from CDS Analyses (Chapter 8) and Output from MCDS Analyses (Chapter 9) in the Users Guide.

You can increase or decrease the font size of an individual results page by right clicking and choosing the appropriate button. This is particularly useful for the text-based cluster size regression plots, which don't fit easily on a page. You can choose Set **current font as default** to make the font size the default for all results and log pages.

For information about transferring your results to another application, see Exporting CDS Results in Chapter 8 of the Users Guide, Exporting MCDS Results in Chapter 9, or Exporting MRDS Results in Chapter 10.

Design Properties Dialog

This dialog window allows you to view and edit the properties for a survey design class. For more information about design classes and survey plans, see the Chapter 6 - Survey Design in Distance of the Users Guide.

The dialog is divided into four tab pages:

- **General Properties**
- **Effort Allocation**
- **Sampler**
- **Coverage Probability**

In addition there are three buttons at the bottom of the page:

- **Defaults** - resets all options on all tab pages to the Distance defaults
- **OK** - saves any changes and closes the dialog
- **Cancel** - closes the dialog without saving changes

The **General Properties** and **Coverage Probability** tabs are identical for all designs, while there is a different tab for each **Sampler** type (point and line). The **Effort Allocation** tab displays a different set of properties for each design class. The set of possible sampler and design class combinations is as follows:

Sampler	Design Class
Point	Simple Random Sampling
	Systematic Grid Sampling
Line	Parallel Random Sampling
	Systematic Random Sampling
	Systematic Segmented Trackline Sampling
	Systematic Segmented Grid Sampling
	Equal Angle Zigzag
	Equal Spaced Zigzag

	Adjusted Angle Zigzag

The statistics from any design class run include the minimum, maximum, mean, and standard deviation of the coverage probability. For a survey plan the statistics from the run include the number of points or lines, the maximum possible area coverage, the realized area coverage, and the mean realized proportion of survey area covered. Each statistic is the sum over all strata. If the design is based on a line sampler then the statistic for the mean realized sampler line length (mean of all strata) is also generated.

General Design Properties Tab

The **General Properties** tab lets you choose the stratum layer, coordinate system and random number generator (RNG) seed value. This tab is the same for all design classes.

The Stratum Layer

Choose the stratum layer from the drop-down list of available stratum layers. The global, stratum and substratum type layers are displayed in the list. The survey design is generated within each region stored in the chosen stratum layer. A description of the stratum layer's coordinate system is given. The stratum layer can be either geo-referenced or not. If it is not geo-referenced, the distance measurement units of its non-earth coordinate system will be displayed. The coordinates of a geo-referenced stratum layer are stored in degrees latitude and longitude and the type of geographic coordinate system is shown. If the stratum layer coordinates are projected, then the description and units of the map projection are also displayed.

The Design Coordinate System

The survey designs are generated within the stratum layer regions whose coordinates correspond to the design coordinate system selected. If the selected stratum layer is stored within a non-earth coordinate system, then the box "Same coordinate system as stratum" is checked, and the Non-earth referenced radio button is selected by default. The reason for this is that the facilities to transform the non-earth coordinates of the stratum layer into a geographic or projected coordinate system are not available. This type of geo-referencing can be achieved in most commercially available GIS packages. For a non-earth stratum layer, the survey design takes place in the same non-earth coordinate system in which the original stratum layer coordinates are stored. If the stratum layer coordinates are stored as degrees latitude and longitude or if they are projected, then a checked "Same coordinate system as stratum" box leads to the designs being generated in the same geographic or projected coordinate system as the stratum layer. Unchecking the box lets you choose a design coordinate system that is different from that of the stratum layer.

If your survey takes place over a small geographic area, the selected geo-coordinate system and map projection are not overly important, and you may well store your stratum layer coordinates within a non-earth coordinate system. However, an appropriate selection is crucial for larger survey regions. The geo-coordinate system and projection chosen will make a difference to the results. Representing the Earth's surface in two dimensions causes distortion in the shape, area, distance, or direction of the data. Different projections cause different types of distortions. Some projections are designed to minimize the distortion of one or two of the data's characteristics. For instance, a projection could maintain the area of a feature but alter its shape.

Equal-area projections preserve the surface area of displayed features, but this is achieved by distorting the shape, angle, and scale properties. As abundance estimation requires an accurate value for the surface area of the study survey region, an appropriately chosen equal-area projection should be used to calculate

the areas. Equidistant projections preserve the distances between certain points, which is an important consideration when calculating the length of line samplers or the distance between sampling locations. True-direction or azimuthal projections maintain the directions or azimuths of all points on the map correctly, while conformal projections preserve local shape, both of which play an important role in navigation. For a more detailed description of the various types of geo-coordinate systems and projections, as well as some guidelines for selecting an appropriate combination of these, see the section on Coordinate Systems and Distance Data (Chapter 5) and Coordinate Systems, Maps and Calculations in Distance (Chapter 5) of the Users Guide.

Generally, the coordinates of a geo-referenced survey region will be stored as decimal degrees latitude and longitude and a projection will be chosen as the design coordinate system. Selecting the Geographic coordinate system radio button will seldom lead to reasonable results. In the rare instance where the design is generated within a geo-coordinate system, it will be the same as that of the stratum layer. Clicking the third radio button will let you select options for the projected design coordinate system. The geo-coordinate system on which the projection is based will be the same as that of the stratum layer. Select the projection and the distance measurement units associated with it. If the stratum is not projected (the most common case), then the design's projection is set to that of the project by default (unless this is [None], in which case the first in the list of projections is set as the default).

The Random Number Generator (RNG)

Pseudo-random numbers are used when running the designs, to provide a "random" starting point for each survey. The numbers are produced by a random number generator (RNG). The RNG uses a "seed" to start the sequence of pseudo-random numbers. By clicking on the "from system clock" radio button the value of the seed is taken from the computer's system clock. By selecting the other radio button the seed value can be set to a fixed value (which must be an odd whole number greater than 2 million). If you set the seed to a fixed value, then each run will produce the same results – useful if you want to generate exactly the same survey again in the future. If you use the system clock, then the actual value used is recorded in the results, so you can generate the same results again if you want.

Coverage Probability Design Properties Tab

The **Coverage Probability** tab lets you choose between an analytic (assume even) or simulation based estimation of the coverage probabilities, at what resolution these estimates should be made by selecting the coverage grid, and where the results should be stored. For more details, see the Users Guide section on Concept: Coverage Probability in Chapter 6.

If you select the first radio button the coverage probabilities will be estimated analytically, and assumed even. The estimates of coverage probability that are calculated analytically are only approximate. A simplistic formula that gives the proportion of the sampled area relative to the survey region surface area is used. It does not take into account sampler overlap, or the fact that parts of the sampler may fall outside the study area along the region boundary. This option will most likely be chosen if you already know about the coverage probability achieved by a given design class, and are interested in its other properties – analytic estimates are *much* faster to calculate than simulated estimates.

To estimate the coverage probabilities by simulation, select the second radio button. You then need to set the number of repetitions of the survey design that should be used to obtain the estimates. For the purpose of estimating coverage probabilities, point or line transects have an associated radius or half-width, respectively. The precision of the coverage probability estimates obtained is highly dependent on the grid point spacing relative to the sampler width or

radius, and on the average proportion of the survey region sampled by each design in the design class. Thus it is important to select an appropriate grid spacing before starting the simulations. A spacing that is too coarse provides a very noisy estimate of coverage probability because for most realizations of this design, a large number of samplers remain undetected by any grid point. A very fine point grid is computationally intensive (i.e. unduly slow). Good grid spacing ensures that a sufficient number of grid points are hit by the sampler, relative to its surface area. A basic rule of thumb is that the number of grid points falling within a single sampler - for any particular realization of the design - should never be less than one at a bare minimum.

Increasing the number of simulations can counteract the shortfall caused by a coarse grid. It is more effective, however, to increase the resolution of the grid. To obtain a sufficiently precise estimate of coverage probability, while not sacrificing computational efficiency, a trade-off between grid spacing and the total number of simulations must be made. We suggest the following approximate formula. If you require a variance no greater than v in your coverage probability estimates, then the total number of simulations needed is approximated by $A^2v / a(A-a)$, where a is the total surface area covered by the samplers and A that of the survey region. This approximation assumes equal coverage probabilities. If the coverage probability is not even, then the number of simulations must be increased to achieve the same precision.

The intrinsic stochasticity in the estimation process means that the coverage probability estimates will be variable for even as well as uneven probability designs. The level of variability decreases as the number of simulations increases, but given the coverage probability results and a case where uncertainty exists about whether the design provides even cover or not, one can test for even coverage using an index-of-dispersion or a classical χ^2-goodness-of-fit test.

Note that points falling in adjacent strata that are hit by portions of the sampler lying outside the design stratum are disregarded. This is justified by the assumption that potential observations at the points in question would not be recorded from the sampler over the stratum boundary.

Results Coverage Grid

The regular grid of points contained in the coverage layers is used for estimating coverage probability for our survey designs. You select the coverage grid layer in which the coverage probabilities are stored from the previously created drop-down list of grid layers. The coverage probability estimates at each point in the coverage probability grid are stored in the field whose name you specify in the text box.

Sampler Design Properties Tab

Sampler tab for point sampler designs

The **Sampler** tab for design classes based on point samplers allows you to specify the radius associated with each point sampler. This radius is required to estimate the coverage probabilities. We suggest you use the value of your truncation distance.

Select the point sampler radius units from the drop-down list. If the design coordinate system is non-earth or projected, these are linear distance measurement units. If the design takes place in a geo-coordinate system, these are angular units. It is best to choose the same units that are used in the design coordinate system and for effort allocation, as then Distance won't have to convert between different units. Conversion inevitably leads to some loss in precision (although this loss is usually very small).

Check the **Same properties for all strata** box if you want the point samplers to have the same radius in all survey strata. The box will be checked and

disabled if there is only a single stratum in the selected stratum layer. Enter a single positive value for the radius in the **Radius** column of the grid table. Unchecking the box will expand the grid table. Each row in the table will correspond to a stratum in the layer, which allows you to enter a different radius value for each stratum (if, for example, you have a different truncation distance in different strata). Each stratum's label, or ID value if the strata are not labelled, are shown in the **Stratum** column of the table.

Sampler tab for line sampler designs

The **Sampler** tab for design classes based on line samplers allows you to specify the half-width associated with each line sampler. The width is required to estimate the coverage probabilities. We suggest you use the value of your truncation distance.

Select the line sampler half-width units from the drop-down list. If the design coordinate system is non-earth or projected, these are linear distance measurement units. If the design takes place in a geo-coordinate system, these are angular units. It is best to choose the same units that are used in the design coordinate system and for effort allocation, as then Distance won't have to convert between different units. Conversion inevitably leads to some loss in precision (although this loss is usually very small).

Check the **Same properties for all strata** box if you want the line samplers to have the same half-width in all survey strata. The box will be checked and disabled if there is only a single stratum in the selected stratum layer. Enter a single positive value for the half-width in the **Width** column of the grid table. Unchecking the box will expand the grid table. Each row in the table will correspond to a stratum in the layer, which allows you to enter a different half-width value for each stratum (if, for example, you have a different truncation distance in different strata). Each stratum's label, or ID value if the strata are not labelled, are shown in the **Stratum** column of the table.

Effort Allocation Design Properties Tab

The Effort Allocation property pages let you define the amount of effort you want to apportion to each stratum in the survey layer.

For survey layers containing multiple strata, you can allocate zero effort to some of the strata. For those strata with zero effort, the design properties will not be calculated during a design run, and no design will be generated during a survey run. The sum of the effort over all strata should, however, be greater than zero.

Simple Random Sampling - Effort Allocation Properties

Edge Sampling

The Edge Sampling options provide different methods for dealing with point samplers falling along the boundary of the survey region. For more information, see the section on Concept: Edge Effects in Chapter 6 of the Users Guide.

Allocation by stratum

Each row in the grid table corresponds to a stratum in the layer, which allows you to allocate effort for each stratum in the survey layer. Each stratum's ID and label (if this field exists) are shown in the **Id** and **Label** column of the table, respectively.

You can select the **Absolute values** radio button and enter the number of point samplers you want in the **Effort** column of the grid table. Otherwise, if you

select the other radio button you can enter a point sampler total in the text box and specify a percentage from that total in the **Effort%** column of the grid table. The percentages over all the strata do not have to sum to 100. Under the second option the **Integer Totals** box will also be enabled. By checking this box any effort percentage that leads to a non-integer number will be rounded to an integer. Point samplers are always generated from integer totals anyway. Check the **Same effort for all strata** box if you want the same number or percentage of point samplers in all survey strata, otherwise you can allocate different effort values for each stratum. The box will be checked and disabled if there is only a single stratum in the selected stratum layer. If the box is checked when the stratum layer contains multiple strata then a single row is displayed in the table, and the effort values entered in this row are used for all strata. The Total points text box displays the aggregated total of sampler points over all survey strata.

Each point sampler is stored as a sampling unit when you create a survey plan.

Systematic Grid Sampling - Effort Allocation Properties

Edge Sampling

The **Edge Sampling** options provide different methods for dealing with point samplers falling along the boundary of the survey region. For more information, see the section on Concept: Edge Effects in Chapter 6 of the Users Guide..

Allocation by stratum

Select the between grid point spacing units from the drop-down list. If the design coordinate system is non-earth or projected these are linear distance measurement units. Otherwise, if the design takes place in a geo-coordinate system these are angular units. By selecting the same units that are used in the design coordinate system or for the sampler radius, imprecision introduced during unit conversions can be avoided.

Each row in the grid table corresponds to a stratum in the layer, which allows you to allocate effort for each stratum in the survey layer. Each stratum's ID and label (if this field exists) are shown in the **Id** and **Label** column of the table, respectively.

You can select the **Absolute values** radio button and enter the number of point samplers you want in the **Effort column** of the grid table. The second radio button lets you enter a point sampler total in the text box and specify a percentage from that total in the **Effort%** column of the grid table. The percentages over all the strata do not have to sum to 100. By selecting the **Systematic point grid spacing** you can enter the regular spacing between grid points. If the **Square grid** box is checked the point grid is square and you enter the length of each square's side under the **Side** column in the table. With this box unchecked you can enter different values horizontal and vertical grid spacing values in the **Width** and **Height** columns, respectively. Under the second and third effort allocation option the **Integer Totals** box will also be enabled. By checking this box any effort percentages or grid spacing that leads to a non-integer number will be rounded to an integer. Point samplers are always generated from integer totals anyway. Enter the angle of the systematic point grid with respect to the x-axis - measured in an anti-clockwise direction from the positive x-axis – in the table's **Angle** column. The angle should be greater or equal to zero and less than ninety degrees.

When the **Update effort in real time** box is checked calculations to estimate the missing information are performed. So, if you enter an absolute number of points the software tries to estimate the square spacing required for a systematic grid with that number of points. In general a systematic spacing can be found that gives an estimated number of points near to – rather than exactly equal to – the absolute number you have specified. This is where the **Effort Tolerance** text box comes in. This allows the calculated spacing to give an estimated number that lies within the effort tolerance percentage number of points either

side of the actual number you have specified. If you specify an effort tolerance that is really narrow, then the software may not be able to find a spacing. If the tolerance is too wide, the algorithm may stop before a potentially better spacing is found. So, try starting off with a fairly narrow tolerance, say one (1%), and if you keep getting error messages make it wider. If you enter a grid spacing, the number of points resulting from this spacing will be estimated. The points are generated according to the spacing specified or estimated for the systematic regular grid of sampler points, so the number of point samplers generated in a run of the design may differ from the absolute number specified or the approximate number calculated on this page – you should therefore use the estimates on this page only as a guide. If you change the distance units then the grid spacing for each stratum is updated as are the estimated number of points for that new spacing. Alternatively, if your computer is slow or you want to enter all your values and then do the calculations just uncheck the **Update effort in real time** box, and press the **Update Effort** button when you are ready.

Check the **Same effort for all strata** box if you want either the same grid spacing, or same number or percentage of point samplers in all survey strata, otherwise you can allocate different values for each stratum. The box will be checked and disabled if there is only a single stratum in the selected stratum layer. The Total points text box displays the estimated aggregated total of sampler points over all survey strata.

Each point sampler is stored as a sampling unit when you create a survey plan. Future versions of Distance may allow you to store this design in two sample layers – points along lines, allowing you to choose the appropriate level of analysis in the analysis engine.

Parallel Random Line Sampling - Effort Allocation Properties

Edge Sampling

The Edge Sampling options provide different methods for dealing with line samplers falling along the boundary of the survey region. For more information, see the section on Concept: Edge Effects in Chapter 6 of the Users Guide.

Effort determined by

Select the first radio button if you want to determine effort by **Sampler number** (i.e., number of lines). With this option the number of survey lines you specify in the **Samplers** column of the table will be generated. If you select the second **Sampler length** option and specify the length value in the **Length** column of the table, then sampler lines will be generated until their aggregated length exceeds the length specified.

Allocation by stratum

Select the line length units from the drop-down list. If the design coordinate system is non-earth or projected these are linear distance measurement units. Otherwise, if the design takes place in a geo-coordinate system these are angular units. By selecting the same units that are used in the design coordinate system or for the sampler width, imprecision introduced during unit conversions can be avoided.

Each row in the grid table corresponds to a stratum in the layer, which allows you to allocate effort for each stratum in the survey layer. Each stratum's ID and label (if this field exists) are shown in the **Id** and **Label** column of the table, respectively.

You can select the **Absolute values** radio button and – depending on which **Effort determined by** option you chose – enter either the number or aggregated length of line samplers you want in the **Samplers** or **Length** column of the grid table, respectively. The second radio button lets you enter a number or aggregated length of line total in the text box, and then specify a

percentage from that total in the **Effort%** column of the grid table. The percentages over all the strata do not have to sum to 100. Under the second effort allocation option the **Integer Totals** box will also be enabled. By checking this box any effort percentages that lead to a non-integer line number will be rounded to an integer. If effort is determined by line number, then these samplers are always generated from integer totals anyway. Enter the angle of the parallel line samplers with respect to the x-axis - measured in an anti-clockwise direction from the positive x-axis – in the table's **Angle** column. The angle should be greater or equal to zero and less than 180 degrees.

When the **Update effort in real time** box is checked calculations to estimate the missing information are performed. So, if effort is determined by **Sampler number** and you enter an absolute number of points or a percentage value, the software tries to estimate the line length of sampler line that would be generated. This is only an approximation, which is dependent on the shape of your survey region. In the current version of the software the approximation may also grow worse as the angle of the sampler lines departs from 90 degrees. If effort is determined by **Sampler length**, then as you enter an absolute line length or a percentage value, the software tries to estimate the number of sampler lines that would be generated. If you change the distance units then the line length for each stratum is updated, as are the estimated number of lines for that new length. Alternatively, if your computer is slow or you want to enter all your values and then do the calculations just uncheck the **Update effort in real time** box, and press the **Update Effort** button when you are ready.

Check the **Same effort for all strata** box if you want either the same line length, the same number of lines, or percentage of sampler number or length, in all survey strata. Otherwise you can allocate different values for each stratum. The box will be checked and disabled if there is only a single stratum in the selected stratum layer. The Total lines and length text boxes display the exact or estimated aggregated totals of sampler lines or line length over all survey strata, respectively.

Each line sampler is stored as a sampling unit when you create a survey plan. A single line sampler may be made up of one or more parts, depending on whether the shape of the survey region causes a split in the line.

Systematic Random Line Sampling - Effort Allocation Properties

Edge Sampling

The Edge Sampling options provide different methods for dealing with line samplers falling along the boundary of the survey region. For more information, see the section on Concept: Edge Effects in Chapter 6 of the Users Guide.

Allocation by stratum

Select the line length and spacing units from the drop-down list. If the design coordinate system is non-earth or projected these are linear distance measurement units. Otherwise, if the design takes place in a geo-coordinate system these are angular units. By selecting the same units that are used in the design coordinate system or for the sampler width, imprecision introduced during unit conversions can be avoided.

Each row in the grid table corresponds to a stratum in the layer, which allows you to allocate effort for each stratum in the survey layer. Each stratum's ID and label (if this field exists) are shown in the **Id** and **Label** column of the table, respectively.

You can select the **Absolute values** radio button, and depending on whether you choose **lines** or **line length** from the drop-down list, you then enter either the number or aggregated length of line samplers you want in the **Samplers** or **Length** column of the grid table, respectively. The second radio button lets you

enter a number or aggregated length of line total – depending on the choice of **lines** or **line length** from the drop-down list - in the text box, and then specify a percentage from that total in the **Effort%** column of the grid table. The percentages over all the strata do not have to sum to 100. Under the second effort allocation option the **Integer Totals** box will also be enabled. By checking this box any effort percentages that lead to a non-integer line number will be rounded to an integer. The systematic lines are generated according to the spacing specified or estimated from the number of lines specified, so the number of line samplers generated may differ from the absolute number specified or the approximate number calculated. The **Integer Totals** box is disabled when the third effort allocation radio button is selected, because the estimated number of sampler lines is always an integer anyway. Enter the angle of the parallel line samplers with respect to the x-axis - measured in an anti-clockwise direction from the positive x-axis – in the table's **Angle** column. The angle should be greater or equal to zero and less than 180 degrees.

When the **Update effort in real time** box is checked calculations to estimate the missing information are performed. So, if effort is determined by **lines** and you enter an absolute number of lines or a percentage value, the software tries to estimate the systematic inter-line spacing and the line length of sampler line that would be generated. This is only an approximation, which is dependent on the shape of your survey region. In the current version of the software the approximation may also grow worse as the angle of the sampler lines departs from 90 degrees. If effort is determined by **line length**, then as you enter an absolute line length or a percentage value, the software tries to estimate the systematic inter-line spacing and number of sampler lines that would be generated. If effort is determined by **Systematic line spacing**, then as you enter the spacing value the software tries to estimate the number of sampler lines that would be generated, their aggregated length, and the associated percentage value. If you change the distance units then either the line length or inter-line spacing – depending on the selected effort allocation option - for each stratum is updated, as are the values in the remaining columns. Alternatively, if your computer is slow or you want to enter all your values and then do the calculations just uncheck the **Update effort in real time** box, and press the **Update Effort** button when you are ready.

Check the **Same effort for all strata** box if you want either the same line length, the same number of lines, or percentage of sampler number or length, in all survey strata. Otherwise you can allocate different values for each stratum. The box will be checked and disabled if there is only a single stratum in the selected stratum layer. The Total lines and length text boxes display the approximate and exact aggregated totals of sampler lines and line length over all survey strata, respectively.

Each line sampler is stored as a sampling unit when you create a survey plan. A single line sampler may be made up of one or more parts, depending on whether the shape of the survey region causes a split in the line.

Systematic Segmented Line Sampling - Effort Allocation Properties

Edge Sampling

The Edge Sampling options provide different methods for dealing with line samplers falling along the boundary of the survey region. For more information, see the section on Concept: Edge Effects in Chapter 6 of the Users Guide.

Non-convex survey regions

The segmented line sampling design is generated by systematically spacing segments along tracklines. The tracklines can be generated within the survey region or its minimum bounding rectangle. If the design is generated within irregular survey regions that have narrow sub-regions this can lead to uneven or,

in the extreme case, some zero coverage probabilities, if complete segments are used (see below). If simulation shows such an effect then check the **Use a minimum bounding rectangle** box to counteract it. However, for irregular survey regions checking this box may lead to survey designs with a less systematic spatial spread throughout the region. If the designs are generated within the minimum bounding rectangle of the survey region, then sampler segments will sometimes necessarily be less than complete, because they are clipped against the original survey region. For this reason the **Allow split sampler segments** box is disabled when the **Use a minimum bounding rectangle** box is checked.

For designs generated within the survey region itself, portions of segments may also lie outside the survey region, where they intersect the boundary. Check the **Allow split sampler segments** box if you want to allow boundary segments to be broken in two. If this box remains unchecked you will always get complete segments, but the cost of this is some irregularity in the inter-segment spacing. If split segments are disallowed, then the design ensures you get complete segments by moving boundary segments along or between the tracklines. If more than half of the length of a boundary segment falls on its original trackline, then the segment is moved 'inwards' until it completely falls within the survey region. Otherwise, it is removed from its original trackline and placed completely on the next trackline in the sequence.

Allocation by stratum

Select the line length and spacing units from the drop-down list. If the design coordinate system is non-earth or projected these are linear distance measurement units. Otherwise, if the design takes place in a geo-coordinate system these are angular units. By selecting the same units that are used in the design coordinate system or for the sampler width, imprecision introduced during unit conversions can be avoided.

Each row in the grid table corresponds to a stratum in the layer, which allows you to allocate effort for each stratum in the survey layer. Each stratum's ID and label (if this field exists) are shown in the **Id** and **Label** column of the table, respectively.

If you select the **Absolute values** radio button, then you choose **sampler segments** or **total length** from the drop-down list. The **sampler segments** option lets you then enter the number of segment samplers you want, in the **Samplers** column of the grid table. The second option lets you enter the aggregated segment length in the **Length** column. The second radio button lets you enter a total segment number or aggregated segment length – depending on the choice of **sampler segments** or **total length** from the drop-down list - in the text box, and then specify a percentage from that total in the **Effort%** column of the grid table. The percentages over all the strata do not have to sum to 100. Under the second effort allocation option the **Integer Totals** box will only be enabled if **sampler segments** option is showing. By checking this box any effort percentages that lead to a non-integer segment number will be rounded to an integer. Choosing the **Systematic line spacing** effort allocation option lets you specify the inter-segment spacing in the **Spacing** column of the table. If the **Same spacing between segments and lines** box is checked then the tracklines along which the segments run are spaced at the same distance as the segments. By un-checking this box you can specify a different inter-trackline spacing in the **Trackline** column of the table. Enter the length of each sampler segment in the table's **Segment** column. The systematic line segments are generated according to the inter-segment and inter-trackline spacing specified, or estimated from either the specified number of segments or their specified total length. Thus, the number of segment samplers generated may differ from the absolute number specified or the approximate number calculated. The **Integer Totals** box is disabled when the third effort allocation radio button is selected, because the estimated number of samplers is always an integer anyway. Enter the angle of the parallel tracklines with respect to the x-

axis - measured in an anti-clockwise direction from the positive x-axis – in the table's **Angle** column. If you wish use the same angle for each realization, then enter a value greater or equal to zero and less than 180 degrees. Alternatively, if you enter a value of -1 then Distance will choose a random angle for each realization.

When the **Update effort in real time** box is checked calculations to estimate the missing information are performed. So, if effort is determined by **sampler segments** and you enter an absolute number of segments or a percentage value, the software tries to estimate the systematic inter-segment spacing (the inter-trackline spacing is the same) and the total length of sampler segments that would be generated. You most enter a value in the **Segment** column for the calculations can proceed. The result of the calculations is only an approximation, which is dependent on the shape of your survey region. In the current version of the software the approximation may also grow worse if the angle of the sampler lines is not 90 degrees. If effort is determined by **segment length**, then as you enter a total segment length or a percentage value, the software tries to estimate the systematic inter-segments (and trackline) spacing and number of sampler segments that would be generated. Again only if the length of a single segment has been entered. If effort is determined by **Systematic line spacing**, then as you enter the spacing value(s) the software tries to estimate the number of sampler segments that would be generated, their aggregated length, and the associated percentage value. If you change the distance units then either the segment length or inter-segment (and trackline) spacing – depending on the selected effort allocation option - for each stratum is updated, as are the values in the remaining columns. Alternatively, if your computer is slow or you want to enter all your values and then do the calculations just uncheck the **Update effort in real time** box, and press the **Update Effort** button when you are ready.

Check the **Same effort for all strata** box if you want either the same total segment length, the same number of segments, or percentage of sampler number or length, in all survey strata. Otherwise you can allocate different values for each stratum. The box will be checked and disabled if there is only a single stratum in the selected stratum layer. The Total lines and length text boxes display the approximate and exact aggregated totals of sampler segment lines and segment line length over all survey strata, respectively.

Each segment sampler is stored as a sampling unit when you run this design (split segments are stored as separate sampler, which may lead to some inappropriately small samplers. This will be dealt with in future versions of the software). This is fine if spacing between lines and samplers is similar (or in the unlikely case that spacing between samplers is greater then between lines). But not if spacing between samplers is small relative to lines. Future versions of Distance will store this design in two sample layers – segments within lines, allowing you to choose the appropriate level of analysis in the analysis engine.

Systematic Segmented Grid Line Sampling - Effort Allocation Properties

The options for this design are the same as for the systematic segmented line sampling design.

Equal Angle Zigzag - Effort Allocation Properties

Non-convex survey region approximated by a

The Non-convex survey region options provide different methods for dealing with non-convex survey regions (see Zigzag Sampling Non-convex Survey Region Options in the Program Reference).

Effort determined by

Select the first radio button if you want to determine effort by **Sampler angle**. With this option the equal zigzag sampler will be generated with the constant angle you specify in the **Angle** column of the table. The constant angle should be greater than zero and less than ninety degrees. If you select the second **Sampler length** option and specify the zigzag's length value in the **Length** column of the table, then the equal angle corresponding to the length specified will be calculated. The result of the calculations is only an approximation, which is dependent on the shape of your survey region. In the current version of the software the approximation may also grow worse if the angle of the sampler lines is not 90 degrees. The length of the zigzag generated at the calculated angle will thus vary to a lesser or greater degree from the length specified.

Design Axis

The design axis options provide different methods for specifying the orientation of the design axis for zigzag samplers (see Zigzag Sampling Design Axis Options in the Program Reference).

Allocation by stratum

Select the line length units from the drop-down list. If the design coordinate system is non-earth or projected these are linear distance measurement units. Otherwise, if the design takes place in a geo-coordinate system these are angular units. By selecting the same units that are used in the design coordinate system or for the sampler width, imprecision introduced during unit conversions can be avoided.

Each row in the grid table corresponds to a stratum in the layer, which allows you to allocate effort for each stratum in the survey layer. Each stratum's ID and label (if this field exists) are shown in the **Id** and **Label** column of the table, respectively.

You can select the **Absolute values** radio button and – depending on which **Effort determined by** option you chose - enter either the sampler angle or length in the **Angle** or **Length** column of the grid table, respectively. The second radio button is only enabled when effort is determined by length, and lets you enter a length of line in the text box. You can then specify a percentage from that total in the **Effort%** column of the grid table. The percentages over all the strata do not have to sum to 100.

When the **Update effort in real time** box is checked calculations to estimate the missing information are performed. So, if effort is determined by **Sampler angle** and you enter an angle value in degrees, the software tries to estimate the line length of the zigzag sampler that would be generated. Similarly, if effort is determined by **Sampler length**, then as you enter an absolute line length or a percentage value, the software tries to estimate the constant angle of the zigzag. If you change the distance units then the line length for each stratum is updated, as are the angle estimates for that new length. Alternatively, if your computer is slow or you want to enter all your values and then do the calculations just uncheck the **Update effort in real time** box, and press the **Update Effort** button when you are ready.

Check the **Same effort for all strata** box if you want either the same line length or zigzag angle, in all survey strata. Otherwise you can allocate different values for each stratum. The box will be checked and disabled if there is only a single stratum in the selected stratum layer. The Total length text box displays the aggregated totals of sampler line length over all survey strata.

The zigzag sampler is made up of line segments (each determined by a change of zigzag direction). These segments are stored as sampling units when you create a survey plan.

Equal Spaced Zigzag - Effort Allocation Properties

Non-convex survey region approximated by a

The Non-convex survey region options provide different methods for dealing with non-convex survey regions (see Zigzag Sampling Non-convex Survey Region Options in the Program Reference).

Effort determined by

Select the first radio button if you want to determine effort by **Sampler spacing**. With this option the equal zigzag sampler will be generated at the spacing you specify in the **Spacing** column of the table. The equal spaced zigzag passes through equally spaced points on opposite sides of the survey region boundary, and this value determines the spacing of those points. If you select the second **Sampler length** option and specify the zigzag's length value in the **Length** column of the table, then the equal spacing corresponding to the length specified will be calculated. The result of the calculations is only an approximation, which is dependent on the shape of your survey region. In the current version of the software the approximation may also grow worse if the angle of the sampler lines is not 90 degrees. The length of the zigzag generated at the calculated spacing will thus vary to a lesser or greater degree from the length specified.

Design Axis

The design axis options provide different methods for specifying the orientation of the design axis for zigzag samplers (see Zigzag Sampling Design Axis Options in the Program Reference).

Allocation by stratum

Select the line length units from the drop-down list. If the design coordinate system is non-earth or projected these are linear distance measurement units. Otherwise, if the design takes place in a geo-coordinate system these are angular units. By selecting the same units that are used in the design coordinate system or for the sampler width, imprecision introduced during unit conversions can be avoided.

Each row in the grid table corresponds to a stratum in the layer, which allows you to allocate effort for each stratum in the survey layer. Each stratum's ID and label (if this field exists) are shown in the **Id** and **Stratum** column of the table, respectively.

You can select the **Absolute values** radio button and – depending on which **Effort determined by** option you chose - enter either the sampler spacing or length in the **Spacing** or **Length** column of the grid table, respectively. The second radio button is only enabled when effort is determined by length, and lets you enter a length of line in the text box. You can then specify a percentage from that total in the **Effort%** column of the grid table. The percentages over all the strata do not have to sum to 100.

When the **Update effort in real time** box is checked calculations to estimate the missing information are performed. So, if effort is determined by **Sampler spacing** and you enter a spacing value, the software tries to estimate the line length of the zigzag sampler that would be generated. Similarly, if effort is determined by **Sampler length**, then as you enter an absolute line length or a percentage value, the software tries to estimate the constant spacing of the zigzag. If you change the distance units then the spacing or line length – depending on how effort is determined - for each stratum is updated, as are length or spacing estimates corresponding to the new spacing or length, respectively. Alternatively, if your computer is slow or you want to enter all your values and then do the calculations just uncheck the **Update effort in real time** box, and press the **Update Effort** button when you are ready.

Check the **Same effort for all strata** box if you want either the same line length or zigzag spacing, in all survey strata. Otherwise you can allocate different values for each stratum. The box will be checked and disabled if there

is only a single stratum in the selected stratum layer. The Total length text box displays the aggregated totals of sampler line length over all survey strata.

The zigzag sampler is made up of line segments (each determined by a change of zigzag direction). These segments are stored as sampling units when you create a survey plan.

Adjusted Angle Zigzag - Effort Allocation Properties

Non-convex survey region approximated by a

The Non-convex survey region options provide different methods for dealing with non-convex survey regions (see Zigzag Sampling Non-convex Survey Region Options in the Program Reference)

Effort determined by

Select the first radio button if you want to determine effort by **Coverage probability**. With this option the adjusted angle zigzag sampler will be generated approximately at the coverage probability you specify in the **Cov Prob** column of the table. Given this value and the line sampler width 4) the length of the zigzag can be determined. Thus, if you do not want the default sampler width, you need to change this before the length calculation takes place. If you select the second **Sampler length** option and specify the zigzag's length value in the **Length** column of the table, then given the sampler width, the coverage probability corresponding to the length specified will be calculated.

Design Axis

The design axis options provide different methods for specifying the orientation of the design axis for zigzag samplers (see Zigzag Sampling Design Axis Options in the Program Reference).

Allocation by stratum

Select the line length units from the drop-down list. If the design coordinate system is non-earth or projected these are linear distance measurement units. Otherwise, if the design takes place in a geo-coordinate system these are angular units. By selecting the same units that are used in the design coordinate system or for the sampler width, imprecision introduced during unit conversions can be avoided.

Each row in the grid table corresponds to a stratum in the layer, which allows you to allocate effort for each stratum in the survey layer. Each stratum's ID and label (if this field exists) are shown in the **Id** and **Label** column of the table, respectively.

You can select the **Absolute values** radio button and – depending on which **Effort determined by** option you chose - enter either the sampler coverage probability or length in the **Cov Prob** or **Length** column of the grid table, respectively. The second radio button is only enabled when effort is determined by length, and lets you enter a length of line in the text box. You can then specify a percentage from that total in the **Effort%** column of the grid table. The percentages over all the strata do not have to sum to 100.

When the **Update effort in real time** box is checked calculations to estimate the missing information are performed. So, if effort is determined by **Coverage probability** and you enter a value, the software tries to estimate the line length of the zigzag sampler that would be generated using this value and the sampler width. Similarly, if effort is determined by **Sampler length**, then as you enter an absolute line length or a percentage value, the software tries to estimate the coverage probability of the zigzag, also using the sampler width. If you change the distance units then the line length for each stratum is updated, as are coverage probability estimates corresponding to the new length, respectively. Alternatively, if your computer is slow or you want to enter all your values and

then do the calculations just uncheck the **Update effort in real time** box, and press the **Update Effort** button when you are ready.

Check the **Same effort for all strata** box if you want either the same line length or zigzag coverage probability, in all survey strata. Otherwise you can allocate different values for each stratum. The box will be checked and disabled if there is only a single stratum in the selected stratum layer. The Total length text box displays the aggregated totals of sampler line length over all survey strata.

The zigzag sampler is made up of line segments (each determined by a change of zigzag direction). These segments are stored as sampling units when you create a survey plan.

Zigzag Sampling Design Axis Options

For zigzag sampling designs, the option you select in the **Design Axis** section of the **Effort Allocation** page determines the angle of the design axis. Zigzag samplers are orientated with respect to this axis. If you select the **Runs at an angle to the x-axis** option, the angle of the design axis is defined with respect to the x-axis - measured in an anti-clockwise direction from the positive x-axis – and can be entered in the "DA Angle" table column on that page. The angle should be greater than or equal to zero and less than 180 degrees. Selecting **Determined by a start and end location** lets you define the start and end locations for the design axis in each stratum. Do this by entering the coordinate values in the **StartX**, **StartY**, **EndX**, and **EndY** table columns. For this third way of defining the design axis, if the stratum layer is geo-referenced, then the **Defined as geographic coordinates** box is enabled. Check the box if you want to enter the coordinates in degrees longitude (x-coordinate) and latitude (y-coordinate).

Zigzag Sampling Non-convex Survey Region Options

This section describes the options on the **Non-convex survey region approximated by a** part of the **Effort allocation** tab for zigzag sampling design classes. For important background information, see Zigzag Sampling Non-convex Survey Region Options in the Program Reference.

Zigzag sampling designs can only be generated in a convex survey region. If any of the survey strata in the survey layer are non-convex then you can choose to generate the design for each stratum in either a **Convex hull** or **Bounding rectangle**, by selecting the appropriate radio button. For non-convex strata the zigzag sampler will no longer be continuous. The amount of discontinuity can generally be reduced by selecting the **Convex hull** option, but this may lead to uneven coverage probabilities. If simulation shows such an effect to be extreme, then select the **Bounding rectangle** option instead.

If you want to store the convex regions within which the designs are generated for some or all of the strata, then check the box and enter a valid name for the new data layer. This new data layer will only be created during survey runs rather than design runs. The convex layer will appear beneath the survey stratum you selected for the design in the data layer hierarchy. If a layer of the same name already exists there, it will be overwritten.

In the current version of the software when you choose the **Bounding rectangle** option to deal with non-convex regions, the bounding rectangle's width runs parallel to the design axis. This gives you a chance to choose a design axis that minimizes discontinuity in the zigzag sampler.

Survey Properties Dialog

The Survey Properties dialog allows you to define the survey methods and where the survey data are located. It is used mostly when setting up survey objects as part of data analysis in Distance – see Working with Surveys during Analysis in Chapter 7 of the Users Guide for more information.

A standard survey object is automatically created by the Setup Project Wizard, if you tell the wizard that you want to setup the project ready for analysis, so it is often not necessary to edit the survey properties, unless you are doing complicated analyses, or require multiple survey objects.

The dialog is accessed by clicking the **Properties…** button on the Survey Details Input tab. It is composed of three tabs:

- Survey methods
- Data layers
- Data fields

In addition there are three buttons at the bottom of the page:

- **Defaults** resets all options on all tab pages to the Distance defaults
- **Ok** saves any changes and closes the dialog
- **Cancel** closes the dialog without saving changes

Before saving any changes, Distance checks to see if the Survey is used by any analyses. If it is, and these analyses have results associated with them, Distance will show the Confirm Change dialog.

Survey Methods Survey Properties Tab

See Survey Properties Dialog for an overview.

Options

- **Type of survey** – can be either a Line Transect, Point Transect or Cue Count survey.
- **Observer configuration** – can be either single or double observer. For more on double observer surveys, see Chapter 10 - Mark Recapture Distance Sampling.
- **Distance measurements** – that is the type of distances that were measured in the field. For line transects this can be perpendicular distances or radial distances together with the angle of the object relative to the trackline. For point transects and cue counts only radial distances are measured.

If the distances were collected in intervals (bins), rather than as exact distances, you should read the Users Guide section on Interval (Binned/Grouped) Data in Chapter 8.

- **Observations** – this is whether recorded observations were of single individuals or clusters of individuals. (See also Clusters of Objects in Chapter 8 of the Users Guide)

- **Sampling fraction** – this option is only here for backward compatibility with previous versions of Distance where sampling fraction was specified as a survey property. We now recommend you specify sampling fraction using a multiplier – see Multipliers in CDS Analysis in Chapter 8 of the Users Guide for details. The appropriate multiplier field can be created during the Setup Project Wizard, or you can create it manually.

Data Layers Survey Properties Tab

See Survey Properties Dialog in the Program Reference for an overview of the Survey Properties dialog.

Here, you specify which data layers relate to this survey. If the survey has been completed, choose the observation layer that contains the survey data. If the survey is planned, but not completed, choose the lowest sample or subsample layer that contains the planned survey effort.

When you choose the lowest data layer, the parent layers are displayed in a table. You must then go on to define the **Data fields** in the next tab.

Data Fields Survey Properties Tab

See Survey Properties Dialog in the Program Reference for an overview of the Survey Properties dialog.

Here, you specify the role of the fields in your survey. You need to fill in the table to tell Distance which fields correspond to the following roles for this survey:

- **Area** – the area of each stratum. If this is set to [None] then density can be estimated, but not abundance.
- **Effort** – the line length for line transects, or number of times each point was visited for point transects. (You can set this to [None] for point transects.)
- **Perp distance** – the perpendicular distance of each observation, if applicable
- **Radial distance** – the radial distance of each observation, if applicable
- **Angle** – the angle of each observation, if applicable
- **Cluster size** – the field containing cluster size, if applicable

Depending on the type of survey, some of these will be greyed out in the table – for example if the survey is a point transect then the Perp distance and Angle fields will be greyed out, as they are not applicable.

If you used Distance 3.5, then this tab plays a similar role to the Modelling Types in Distance 3.5. The current setup is more flexible, however, because you can define multiple surveys within a project.

Data Filter Properties Dialog

The Data Filter Properties dialog allows you to view and edit the properties for a data filter. For more information about data filters, and how they are used in analyses, see the pages in Chapter 7 of the Users Guide on Analysis Components.

The dialog is divided into four tab pages:

- Data Selection
- Intervals
- Truncation
- Units

In addition there are three buttons at the bottom of the page:

- **Defaults** resets all options on all tab pages to the Distance defaults
- **Ok** saves any changes and closes the dialog
- **Cancel** closes the dialog without saving changes

Before saving any changes, Distance checks to see if the Data Filter is used by any other analyses. If it is, and these analyses have results associated with them, Distance will show the Confirm Change dialog.

You can change the name of the Data Filter by typing a new name into the Name text box. This name is saved when you press the **Ok** button.

Data Selection Tab

See Data Filter Properties Dialog of the Program reference for an overview of the Data Filter Properties dialog.

This tab page allows you to select a subset of your data for analysis. This feature, when combined with the ability to add extra fields to the dataset, is extremely powerful because it effectively lets you keep different subsets of your study data (species, years, etc) in the same Distance Project. You could, for example, define a separate Data Filter to select each subset separately, but in your analyses use the same set of Model Definitions. Other possible examples of the use of data selection is given in Chapter 8 of the Users Guide within the sections entitled: Stratification and Post Stratification and Multipliers in CDS Analysis.

To activate the data selection criteria options, click on the **+** button and choose the data layer type that you want to define criteria for.

You choose from data layer types, rather than layer names at this stage because it is only when running the analysis that Distance combines your data filter with a survey object to find out which data layers are to be used in that run.

You then type the selection criteria in the space to the right of the layer type.

If you need more while editing a edit or view a long selection criterion, click on the line you want to see more of and press SHIFT-F1 (i.e., the shift and F1 keys) to open the Data Selection Zoom Dialog.

Selection criteria format

It is important to get the selection criterion exactly right, otherwise the analysis will not work. The format is:

```
FieldName Operator Value
```

- `FieldName` is the name of the field that the criterion applies to. If the field contains any spaces or punctuation then put it inside square brackets – e.g., `Cluster size` becomes `[Cluster size]`

- `Operator` is the type of logical operator to perform – valid operators include =, <>, >, >=, <=, IN, LIKE, BETWEEN
- `Value` is the values the operator applies to – for example `20` (a number) or `'1'` (a text value – put it in quotes)

Aside! If you're a database geek, you'll recognize the format as it is the same as that in a SQL WHERE statement.

For example, in the Stratify example project, if we wanted to select only observations with cluster size 1, we would define a criterion on the observation data layer:

```
[Cluster size] = 1
```

Tip! The IN operator is a useful way of selecting on multiple values, e.g., to select cluster sizes 1, 2 and 3, you could say:

```
[Cluster size] IN (1,2,3)
Another useful operator is BETWEEN, as in
[Cluster size] BETWEEN 1 AND 3
```

Selection on multiple data layers

You can define criteria on different layers – for example to select cluster size 1 but only from the Ideal habitat stratum, you would define a second criterion (by clicking the **+** button again), and this time choose the stratum layer type. The second criterion would be

```
Label = 'Ideal habitat'
```

Complex selection criteria

You can join selection criteria on the same layer using up to 40 logical operators `AND`, `OR` and `NOT`, together with brackets if necessary. Examples of more complex criteria on some fictional project might be:

```
[Cluster size] >= 3 AND Beaufort IN (0,1,2,3)
```

```
Species = 'SOSP' AND (Observer = 'LT' OR Observer = 'KBV')
```

You can also manipulate the data using simple functions, such as:

- string functions LEFT, RIGHT, MID
- numerical functions INT, ROUND

For example:

```
LEFT (Observer,1) = 'L'
```

```
INT(Distance)=0
```

Unfortunately, we haven't been able to find a complete reference for which functions are and are not allowed – we will update this section when we do!

Intervals Tab

See Data Filter Properties Dialog of the Program reference for an overview of the Data Filter Properties dialog.

On this page, you specify whether you want your distance data analyzed as interval data (as opposed to exact measurements).

You would use this option under two circumstances:

- Your data were collected in intervals. In this case you would set the intervals here in the default Data Filter and leave them the same for

all analyses. You should read the Users Guide section on Interval (Binned/Grouped) Data in Chapter 8, for more information on dealing with interval data within Distance.

- Your data were collected as exact measurements, but you wish to analyze them as interval data.

Do not use this option if you want to analyze the data as exact, but want to specify intervals for the goodness-of-fit tests. You do this in the Diagnostics page, under the **Detection Function** tab the Model Definition Properties.

To specify intervals for the analysis:

- Click on the "Transform distance data into intervals for analysis" check box.
- Choose the number of intervals.
- If your intervals are evenly spaced, you can choose the Automatic equal intervals option. You then only need to type in the lowest and highest cutpoints. Distance will fill in the others automatically.
- If your intervals are not evenly spaced, choose the Manual and type in the interval cutpoints.

When you select interval data on this tab page, your truncation options change on the Truncation tab page. By default, the data are truncated at the upper and lower cutpoints you have selected. See the Truncation Tab page of the Program Reference for more details.

Selecting the Automatic equal intervals option will give you less flexibility in choosing truncation distances in the Truncation tab. So, even if you have evenly spaced cutpoints, it is often better to use the Automatic equal intervals option to speed entering the cutpoints (this way you only have to type in the lowest and highest cutpoints), but to select the Manual option before going on to the Truncation tab page.

The Automatic equal intervals option is essential when you have a large number of intervals. This is because Distance stores Manual cutpoints in a list, and this list can not be more than 180 characters long. (If you try and enter Manual cutpoints and the list becomes too long, Distance will give you a warning message.) For Automatic cutpoints, Distance only has to store the highest and lowest cutpoint, and number of intervals - so the list is much shorter. The maximum number of cutpoints you are allowed in either case is 30.

What happens on the cutpoint boundaries? Observations that are exactly on the lowest cutpoint boundary are included in the lowest interval. Thereafter, observations that are exactly on a boundary are put in the lower interval band.

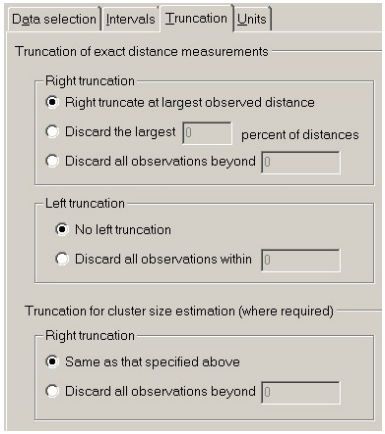
Truncation Tab

See Data Filter Properties Dialog of the Program reference for an overview of the Data Filter Properties dialog.

The Truncation tab is where you choose the level of truncation for your distance data.

Exact distances

This is how the tab looks if you have exact distances i.e., you did not choose to transform your data into intervals on the Intervals tab:



Data Filter Truncation tab when using exact distances

When analyzing distance data, it is common practice to use Right Truncation to remove observations a long way from the track line. In Distance, you can right truncate by discarding a certain percentage of the observations with the largest distances or you can choose a fixed distance for the truncation. We recommend experimenting with running analyses at variety of truncation distances and examining the resulting fit in the Detection Probability Plot and Chi-square Goodness of Fit pages of the Analysis Details Results.

Note! In Distance 3.5 and earlier, when truncating a percentage of data in analyses where the detection function was stratified, the truncation point was calculated separately for each stratum. It is now calculated using all the data, and the same truncation point is applied in all strata. This is more consistent with the other truncation options.

Left truncation (i.e., discarding observations at less than a given distance) is less common. In Distance you can choose a fixed distance for the left truncation.

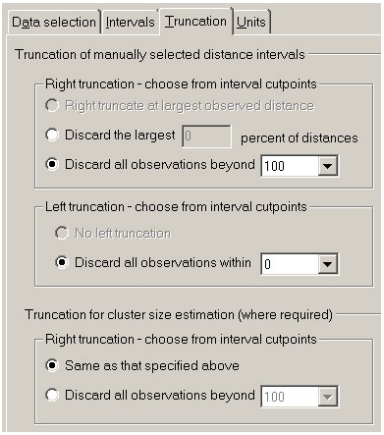
Tip! Left truncation is usually used in CDS when detection on the line is not certain, but detection probability is higher some distance away from the line. Examples include the case where animals close to the line are freeze or hide so are likely to be missed, or aerial surveys where it is hard to see right under the plane. When you use Distance to left truncate, only data beyond the left truncation distance is used, so the detection function is fit only to these data and is extrapolated back to distance zero. The estimated detection probability at the left truncation distance is often less than 1. An alternative analysis method is appropriate when you are willing to assume that detection probability is 1 at the left truncation distance. In this case, you can simply subtract the left truncation distance from the observed distances before importing the data into Distance – effectively moving zero distance out to your left truncation line. In this case, there is no need to specify left truncation within distance – you've already done it before importing the data.

If your observations are clusters, rather than individual animals, you may want to choose a different truncation distance for estimating cluster size. This is especially true if you are running an analysis using the Mean of observed clusters option in the Cluster size tab of the Model Definition Properties. If your observations are not clusters of objects, the cluster size truncation options will be greyed out (see picture).

Note! The cluster size options are only relevant for CDS analyses, and MCDS analyses where cluster size is not a covariate.

Interval data - Manual intervals

When you have specified manual Intervals on the Intervals tab page, the Truncation tab will look something like this:



Data Filter Truncation Tab when data have been transformed into intervals in the Intervals tab

By default, Distance will right truncate all observations that fall beyond your upper interval cutpoint and left truncate all observations that fall within your lower interval cutpoint. If you want to truncate further, you can right or left truncate at any of the interval cutpoints by selecting from the drop-down lists.

For cluster sizes, Distance allows you to choose either the same truncation as above, or to choose from one of your cutpoints.

Interval data - Automatic intervals

If you have specified automatic intervals in the Intervals tab page, then the right and left truncation are set to the upper and lower cutpoints that you specified and cannot be changed.

Units Tab

See Data Filter Properties Dialog of the Program reference for an overview of the Data Filter Properties dialog.

Using the Units tab, you can convert between that the data are stored in (as specified in the Data Explorer), and the units for reporting analysis results. You can, if you want, report results in one unit of area (say) in one analysis and a different set of units in another.

Model Definition Properties Dialog

The Model Definition Properties dialog allows you to view and edit the properties for a model definition. For more information about model definitions

and how they are used in analyses, see the pages in Chapter 7 of theUsers Guide on Analysis Components.

At the top of the dialog, you choose the analysis engine – either CDS (conventional distance sampling – see Users Guide, Chapter 8 - Conventional Distance Sampling Analysis), MCDS (multiple covariates distance sampling – see Users Guide, Chapter 9 - Multiple Covariates Distance Sampling Analysis) or MRDS (mark-recapture distance sampling – see Users Guide, Chapter 10 - Mark Recapture Distance Sampling).

The contents of the dialog change depending on the analysis engine chosen. The CDS and MCDS analysis engines have very similar options, so are grouped together in the pages that follow. The MRDS engine has somewhat different options so is dealt with separately below.

There are three buttons at the bottom of the dialog window:

- **Defaults** resets all options on all tab pages to the Distance defaults
- **Ok** saves any changes and closes the dialog
- **Cance**l closes the dialog without saving changes

Before saving any changes, Distance checks to see if the Model Definition is used by any other analyses. If it is, and these analyses have results associated with them, Distance will show the Confirm Change dialog.

You can change the name of the Model Definition by typing a new name into the Name text box. This name is saved when you press the **Ok** button.

Model Definition Properties - CDS and MCDS

Both the CDS and MCDS analysis engine have the same Model Definition properties tabs:

- Estimate
- Detection Function
- Cluster Size
- Multipliers
- Variance
- Misc.

In the following pages, we discuss the contents of these tabs, highlighting any differences between the engines.

Estimate Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

On the Estimate page you define the stratum and sample layer to use in the analysis, and tell Distance which quantities to estimate. The various options are outlined here, and are also discussed in Chapter 8 of the Users Guide in the section on Stratification and Post-stratification.

Stratum definition

Here, you specify the level of stratification to use in the analysis. For details, see Stratification and Post-stratification in Chapter 8 of the Users Guide.

You specify layer types rather than layer names at this stage because it is only when the analysis is run that the survey object is used to select the data

layers for the analysis. If you select a field for post-stratification that is not in the layers used in the analysis, an error will result.

Sample definition

Here, you specify which sample or sub-sample layer to use as the sample, for determining encounter rate variation and for bootstrapping. For more information see Chapter 8 of the Users Guide, the section on Sample Definition in CDS Analysis.

Quantities to estimate and level of resolution

These options define which quantities you wish to estimate, and at what level. If you have selected **No stratification** in the Stratum definition section then the Stratum column will be greyed out. Also, if your observations are individual objects, not clusters, then the **Cluster Size** row will be greyed out. Lastly, if you are doing an MCDS analysis, and have cluster size as a covariate, the options here will look different (see below).

If you wish to estimate density, you should first check the boxes at the levels for which you want density estimates. In most cases you will not have enough observations in each sample to estimate density by sample, but this is not always true. The lowest level of density dictates the level of estimation for encounter rate, and the lowest level for estimation of the detection function parameters and cluster size. After you have selected the level to estimate density, you can then select one level to estimate the detection function and cluster size (if applicable).

In exploratory analyses, you may not wish to estimate density. For example, you may only be interested in investigating the fit of various detection functions at the global level, and not be interested in estimating density until you have found a satisfactory fit. In this case, uncheck all of the density, encounter rate and cluster size boxes, like this:

	Level of resolution of estimates		
	Global	Stratum	Sample
Density	☐	☐	☐
Encounter rate	☐	☐	☐
Detection function	☑	☐	☐
Cluster size	☐	☐	☐

Notice that the restrictions on the level of resolution of estimates are removed when you are not estimating density.

If you are estimating density by stratum and also globally, you need to tell Distance how to combine the stratum estimates together to make the global estimate. For geographic strata, use the default settings:

If your strata are not spatial strata, for example time periods or different sections of the population, then you should consider using the other options here; however we recommend entering non-spatial data as columns in the appropriate data layer and then using the Post-stratification feature to do the analysis (Chapter 8 of the Users Guide in the section on Stratification and Post-stratification has more information and further details of the scenarios when each of the following options are appropriate).

For reference, the four possible options are:

- Global density estimate is **Mean** of stratum estimates, weighted by **Stratum area**.

This option is appropriate in the usual case where the strata are geographic strata. Here density is calculated as a weighted mean of the stratum estimates, weighted by the area of each stratum. Abundance is calculated as density multiplied by the sum of the stratum areas. Variance is a weighted mean of the stratum variances, although the exact formula depends upon which components (detection function, encounter rate, cluster size) were estimated by stratum and which globally - see Buckland et al. 2001 section 3.7.1.

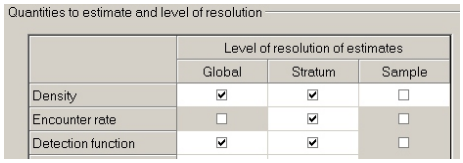
- Global density estimate is **Mean** of stratum estimates, weighted by **Total effort in stratum**.
  This option is appropriate when the strata are effort related, such as different observers or time periods. Here density is calculated as a weighted mean of the stratum estimates, weighted by the sum of the survey effort in each stratum. Abundance is calculated as density multiplied by the area of the first stratum (in almost all cases the strata represent the same area – for example when the strata are different observers surveying the same survey region). In this case the **Strata are replicates** tick box is enabled. The formula for variance depends if this box is ticked
  - **Strata are replicates** ticked.
    Here, strata are seen as a random sample from a larger population of possible strata. An example would be if the strata represent survey days chosen at random (or systematically) over a year, and the inference is about the average density of animals over the whole year. Variance is calculated using the variation in density between strata (weighted by effort); see equations 3.84 - 3.87 in Buckland et al. 2001, treating stratum as a sample.
  - **Strata are replicates** not ticked.
    Here inferences are restricted only to the strata surveyed. An example would be if the strata represent two survey vessels that surveyed the same area and we wish to make inferences about mean density in the area over the two vessels. Here variance is calculated as an effort-weighted average of the stratum variances, using the methods outlined in section 3.7.1 of Buckland et al. 2001, substituting total line length for the area weighting terms (Av and A).
- Global density estimate is **Sum** of stratum estimates.
  This option is appropriate when the strata represent different components of the population, such as male and female animals, and we want a combined estimate of overall density. The stratum density estimates are summed across strata, abundance is the density multiplied by the area of the first stratum (since all strata should have the same area), and the variance is calculated as in Section 3.7.1 of Buckland et al. 2001, dropping the area weighting terms (Av and A).

MCDS analyses

For MCDS analyses, it is possible to fit the detection function at one level and estimate at a lower level. For example you can fit a global detection function model, but estimate average f(0) and probability of detection separately for each stratum. For details of why this may be useful, see Estimating the detection function at multiple levels in Chapter 9 of the Users Guide.

To implement this, you simply check the Levels to estimate boxes for detection function at both the level you want to fit the model (the higher level) and the level you want to estimate average f(0) (the lower level).

For example, imagine you want to estimate density by stratum, but don't have enough data to fit a separate detection function for each stratum. One solution is to fit a global model for the detection function, but using stratum (or lower) level covariates. You can then use the fitted model to estimate a separate average f(0) (or h(0)) in each stratum, using the covariates that apply to that stratum. To implement this, you define the appropriate covariates in the **Covariates** page, and set up the **Quantities to estimate** on the **Estimate** page as follows:



Estimating detection function at multiple levels. Note that the detection function boxes are checked both at the global and stratum levels.

MCDS analyses with cluster size as a covariate

When cluster size is a covariate, the options on this page change:

- **Stratum definition.** The engine currently does not allow for stratification when cluster size is a covariate, so the stratification options are disabled.
- **Sample definition.** Because variance is not estimated analytically, this section is now only used when bootstrapping to estimate variance.
- **Quantities to estimate.** Variance of encounter rate is no longer estimated, so this line is removed, and cluster size is now for output only.

For more information see Cluster size as a covariate in Chapter 9 of the Users Guide.

Detection Function Tab - CDS and MCDS

The Detection Function tab is divided into five pages:

- **Models** - Detection Function Tab - CDS and MCDS
- **Adjustment terms** - Detection Function Tab - CDS and MCDS
- **Covariates** - Detection Function Tab - MCDS (only in the MCDS engine)
- **Constraints** - Detection Function Tab - CDS and MCDS
- **Diagnostics** - Detection Function Tab - CDS and MCDS

Models - Detection Function Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

The Models page is the first that comes under the Detection Function tab. It is probably the most important of all the pages in the Model Definition because it is where you tell Distance what model to use to estimate the detection function.

In almost all cases you will have one detection function model. Select from the drop down list to choose the Key function (Hazard-rate, Half-normal, Uniform or Exponential for CDS analyses, Hazard-rate or Half-normal for MCDS

analyses) and Series expansion (Cosine, Simple polynomial or Hermite polynomial).

To choose a key function with no series expansion, you will need to use manual adjustment in the Adjustment terms page and select number of adjustment parameters to be 0.

The only situation where we recommend you select more than one detection function model is where you are using bootstrapping to incorporate model selection uncertainty in your variance estimate (see Model Averaging in CDS Analysis, in Chapter 8 of the Users Guide). In this case, use the **+** button to add more detection function models, and select from the options under **Selection among multiple models using**. You can choose either AIC, AICc or BIC.

Users of Distance 3.0 and earlier will be used to regularly selecting more than one model in the ESTIMATE section of the command file, and using AIC to select among them. In the current software, we recommend you assign each model to a separate analysis, and then compare the models in the Analysis Browser. This way you have access to Analysis Details for each of the models for example you can look at the Detection Probability Plots in the Results section of the Analysis details.

Adjustment Terms - Detection Function Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

This Adjustment terms page is the second that comes under the Detection Function tab. This is where you tell Distance what methods to use to fit the series expansion terms (Adjustment terms) to the data in the detection function modeling. You can also specify starting values for the estimation procedure, for both the key and adjustment function parameters, and the method for scaling distances when calculating the adjustment terms..

Selection of adjustment terms

Automated selection

Click the Automated selection option for automated selection of adjustment terms. Distance then uses data-based criteria to choose the number and order of adjustment terms required for the analysis.

The selection methods are

- **All** This option examines all possible combinations of a limited number of adjustment terms. If z is the maximum number of parameters and k is the number of parameters in the key function, then there are 2 to the power of z-k combinations of the adjustment terms. Each model is fitted to the data and the model with the smallest value according to the selection criterion is selected.
- **Sequential selection** This option considers a subset of models with different combinations of adjustment terms. A sequence of models is considered, with one adjustment term being added at each step of the sequence. The sequence of models can be represented as
  M1 key function with no adjustment terms
  M2 key function with 1 adjustment term
  M3 key function with 2 adjustment terms

…
Mv key function with v-1 adjustment terms
The selection criterion (stopping rule) is either based on likelihood ratio test or minimizing AIC, AICc or BIC. Model Mt is chosen if there is no model in the sequence Mt+1, ..., Mt+k which provides a significantly better fit as determined by the stopping rule. The value in the Look-ahead field determines the length(k) of the sequence of models that is examined before choosing model Mt. Adjustment terms are added sequentially based on the order of the term. For polynomial adjustment functions, the order of the adjustment term is the exponent of the polynomial. Terms are added in the following sequence: (xt, xt+2,...). For cosine adjustments, cosine terms are added in the following sequence: ($\cos(t\pi x_s)$, $\cos((t+1)\pi x_s)$,...). The beginning value, t, is determined by the shape of the key function. x_s, the scaled perpendicular distance, is defined at the bottom of this page.

- **Forward selection** Forward selection only differs from sequential selection in the choice of which adjustment term is included at each step in the sequence. Forward selection adds one term at a time, but not necessarily in sequential order. For each model in the sequence, each term not already in the model is added and the adjustment term which increases the likelihood the most is chosen as the term to add. With forward selection it is possible to select models that cannot be selected with sequential selection. For example, the following model might be chosen with forward selection:
$f(x) = (1/w)(1+ a1\cos(\pi x_s)+ a2\ \cos(3\pi x_s))$
However, with sequential selection, the adjustment term $\cos(3\pi x_s)$, could not be added without first adding the adjustment term $\cos(2\pi x_s)$.

Manual Selection

Choose this option if you want to specify the number and optionally the order of the adjustment terms yourself. The number of the models on in the tables on this page correspond to the models chosen on the previous **Models** page (normally just 1).

When you type in the number of adjustment parameters, the appropriate number of cells under "Order of adjustment parameters" become editable. Leave these cells blank if you want to specify the number of adjustment parameters but not the order.

To specify a detection function model with just a key function and no adjustment parameters, set the number of adjustment parameters to be 0.

You can also select starting values for the key function and adjustment terms. To do this, check the box **Manually select starting values**, and enter the number of parameters for each model.

Calculate the number of parameters by summing the number of key function parameters, the adjustment terms and any covariate parameters. If the detection function is fit by stratum or sample, sum the number of parameters in each stratum to get the total. For more information about model parameterization, see About CDS Detection Function Formulae in Chapter 8 of the Users Guide.

In newer versions of Distance, you specify the starting values separately for each stratum. In versions 3.5 and earlier, the same starting values were used in each stratum.

Tip! If you are not sure how many parameters there are in an analysis, run it first without specifying starting values, and then look in the Parameter Estimates page of the Analysis Details Results to see how many parameters were used. This is particularly useful for MCDS analyses, as the number of covariate levels in non-factor covariates may vary among strata, depending on which covariate levels occur in each stratum.

Scaling of distances

As explained in Chapter 9, this option is mainly of interest when using the MCDS engine (although it can be used for the CDS engine too). For all of the series expansion terms, the scaled distance is used in place of actual distance in calculating the expansion term values mainly for numerical reasons. There are two options: scale by w, the truncation distance, or by σ, the scale parameter of the key function (this does not apply to the uniform key function, which has no scale parameter). For the CDS engine one will generally want to scale by w, but for the MCDS engine one may scale by either – see Chapter 9 - Multiple Covariates Distance Sampling Analysis in the Users Guide.

Covariates - Detection Function Tab - MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

The Covariates page is the third under the Detection Function tab. This is where you specify the covariates to add to the model. This page only appears for the MCDS analysis engine.

Click on the **+** button to add a row to the table.

In the first column you select the **layer type** containing the covariate from the drop-down list.

Note! This column gives layer types rather than layer names because at this stage Distance doesn't know which Survey you're going to use with this Model Definition, so it doesn't know which layer names it can use. This way you can pair the same Model Definition with many different surveys.

In the second column you select the **field name** of the covariate.

Tick the box in the third column if the covariate is a **factor**. Factor, or class covariates have a finite number of distinct levels. The value of each level is not significant – for example the factor levels could be text fields "Porpoise", "Whale", "Seal", or they could be numeric fields 1, 2, 3. If the box in this column is not ticked, the covariate is assumed to be a **non-factor** covariate. In this case, the field must be numeric, otherwise an error will occur when you try to run the analysis engine. For more about this, see Factor and Non-factor covariates in MCDS in the Users Guide.

Tick the box in the last column if the covariate is the **cluster size** field. Distance needs to know whether any of the fields in your analysis are the cluster size field because it treats this field a special way in the analysis. For more information see Cluster size as a covariate in Chapter 9 of the Users Guide.

Some general advice about selecting covariates to include is given in the CDS Analysis Guidelines section of Chapter 8 - Conventional Distance Sampling Analysis in the Users Guide.

Constraints - Detection Function Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

The Constraints page comes under the Detection Function tab. This is where you specify constraints on the fitting procedure. In most cases, it is sufficient to use the default settings. Other situations are discussed below

Constraints on the shape of the fitted function

This option allows you to choose the level of constraints on the shape of the fitted detection function. The estimators are constrained by default to be strictly monotonically non-increasing – i.e., the detection curve is either flat or decreasing as distance increases from 0 to the truncation distance. In some instances, depending on the tail of the distribution, this can cause a poor fit at the origin (distance = 0), or can cause lack of convergence.

In these cases there are two options: (1) use the data filter to truncate the observations in the tail, or (2) relax the constraints. The weak constraint option allows the curve to go up and down as it fits the data but will not let the curve dip down at the origin. In some instances this will achieve a better fit at the origin, which is the point of interest. With no constraints, the curve can take any form, except it must remain non-negative.

Monotonicity is achieved by constraining the function at a fixed set of points. In some circumstances it is possible that the curve can be non-monotone between the fixed points. Typically, this results from trying to over-fit the data with too many adjustments with a long-tailed distribution. In this case, use a Data Filter to truncate the data, or constrain the number of adjustment terms in the Adjustment Terms tab.

You cannot apply constraints in the MCDS engine. This is because it uses a different fitting algorithm, one for which we have not implemented constraints on the maximization routine.

Bounds on Key Function Parameters

With some datasets, it may be necessary to bound the parameter estimates to achieve convergence. In these situations, you can use this table to specify upper and lower bounds on the key function parameters (one parameter for half-normal and negative exponential key functions, two for the hazard rate function, plus any covariate parameters). One common circumstance where this is required is to impose a lower bound of 1.0 on the second hazard rate parameter.

In newer versions of Distance, you specify bounds separately for each stratum. In versions 3.5 and earlier, the same starting values were used in each stratum. To calculate the number of key function parameters, sum across strata (including any covariates). For more information about model parameterization, see About CDS Detection Function Formulae in Chapter 8 of the Users Guide.

If you are not sure how many key function parameters there are in an analysis, run it first without specifying starting values, and then look in the Parameter Estimates page of the Analysis Details Results to see how many parameters were used. This is particularly useful for MCDS analyses, as the number of covariate levels in non-factor covariates may vary among strata, depending on which covariate levels occur in each stratum.

Diagnostics - Detection Function Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

This page is the last under the Detection Function tab. On this page you specify the diagnostic output that you want Distance to produce, including the intervals for detection probability plots and goodness-of-fit tests, whether to output qq-plots and associated statistics for non-interval data, and the filename if you want Distance to save the plot file.

Chi-sq goodness-of-fit tests and histogram intervals

Note! This option is only relevant if you are analyzing the data as exact distances. If you to analyze your data as intervals (by selecting Intervals in the Data Filter) then a single goodness-of-fit test is performed using those intervals. Any options you set under Intervals on this tab page are ignored.

Automatic selection of intervals

By default, when the data are analyzed as ungrouped, three sets of intervals are constructed with equally spaced cutpoints and the number of intervals being m, 2/3m and 3/2m, where m is the square root of the number of observations.

Manual selection of intervals

Instead of using the automatically generated intervals, we recommend that you routinely select your own. To do this, choose **Manual selection of intervals**, and enter the number of intervals in the text box provided. You then have two choices to specify the interval cutpoints:

- **Automatic equal intervals**. If you choose this option, and you leave the first and last cutpoint values in the **Cutpoints** table as 0, then Distance will generate goodness-of-fit tables with evenly spaced cutpoints between the left truncation distance and the right truncation distance. Alternatively, you can specify the first and last cutpoints in the **Cutpoints** table, and Distance will generate evenly spaced cutpoints between these limits.
- **Manual**. This option allows you to specify uneven cutpoints – for example you may want smaller intervals close to zero distance (to look for evidence of a shoulder) or perhaps cutpoints that isolate favoured distances (to look for evidence of rounding). Enter the cutpoint values you want in the **Cutpoints** table. You will normally set the first cutpoint to the left truncation distance, and the last one to the right truncation distance.

Tip! A good way to get to know your data when you begin analyzing it is to define a large number of intervals (say 15-20), fit any arbitary model, and then examine the output histogram for evidence of evasive movement, heaping, outliers, etc. (you can ignore the model fit for now). See CDS Analysis Guidelines in Chapter 8 of the Users Guide for more on this.

Tip! The **Interval cutpoints** options can make it easier for you to enter manual intervals. For example, you can enter the left and right truncation points in the first and last cutponts rows, and then click on **Automatic equal intervals** to have the intermediate cutpoints set. Then go back to **Manual** and customize the cutponts to your requirements.

K-S test goodness-of-fit test and qq plots

Note! This option is only relevant if you are analyzing the data as exact distances.

Qq plots are a graphical technique for assessing the adequacy of the fit, and the associated test statistics (Kolmogorov-Smirnov and Cramér-von Mises) test goodness-of-fit for exact data. For more about these outputs, see CDS Qq-plots and CDS Goodness of fit tests in Chapter 8 of the Users Guide.

Since these outputs can take a while to generate for large datasets, there is an option here to turn them off. Qq plots have one plotted point per observation, so for large datasets it is better to plot only a subset of points. By default, the maximum number of points to plot is 500, but that can be changed here. Entering 0 under **Maximum num points in qq plots** means that all points are plotted, regardless of how many there are.

Plot file

If higher quality graphical output is required, Distance can save the histogram data to a file that can then be imported into any graphics or statistics package. Check the Create file of histogram data option and choose the file name using the Browse button.

The output format of the file is described in the Users Guide page Saving CDS results to file in Chapter 8.

Tip! You can copy and paste the high quality plots produced by Distance straight into most word processor and spreadsheet packages. In addition, you can easily paste the plot data into a spreadsheet and re-create the plot that way. See the Analysis Details Results Tab help page.

Cluster Size Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

The cluster size page is used to modify the way the cluster sizes are used in the estimate of density.

By default, distance uses a regression of observed cluster size against distance to estimate the average population cluster size (as the expected value at distance of 0). This approach is intended to reduce bias if there is a tendency for smaller clusters to be missed more than large clusters at large distances from the track line. The exact regression method is chosen by the option in the Size-bias regression method box.

A second alternative if cluster size bias is not anticipated to be a problem is to use the mean of the observed clusters as an estimate of the average cluster size in the population. Truncation becomes particularly important if you are using this option (see note below).

Thirdly, you can apply a statistical hypothesis test to the regression of cluster size on distance and only apply the size bias method if this regression is statistically significant.

Fourthly, in MCDS analyses, you can use cluster size as a covariate in the detection function. Doing this makes the options on this page un-necessary, and the **Cluster size** tab is therefore disabled. For more information see the section in Chapter 9 of the Users Guide entitled: Cluster size as a covariate.

Note! Another important cluster size option is the truncation of distances for cluster size calculations. This is set in the Truncation tab of the Data Filter Properties window.

Multipliers Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

In this page you tell Distance which multipliers to use to scale the density estimate. Multipliers are discussed in more detail in Chapter 8 of the Users Guide under Multipliers in CDS Analysis.

To add a new multiplier to the list, press the **+** button. The column "Layer type containing multiplier" is set to the Global data layer and cannot be changed in this version of Distance. You should choose the multiplier you want from the list of multiplier fields that pops up when you click on the "Fields containing multiplier value" column. If this multiplier has a standard error associated with it, choose the appropriate Multiplier SE field from the list that pops up when you click on the "Field containing multiplier SE" column. Note that you don't have to select an SE if your multiplier value is known without error. Similarly, if you know the degrees of freedom (DF) associated with the multiplier, click on the "Field containing multiplier DF" column. You don't have to select a DF; if you do not select one then Distance assumes the DF for this multiplier is infinity. (Another way to tell Distance that DF is infinity is to select a field from the global layer that has a value of 0.0.) Lastly, you must tell Distance what operator to use: i.e., whether to divide or multiply the density estimate by the multiplier value to obtain the final estimate.

To remove the most recently added multiplier press the **-** button.

Note! If the multiplier represents cue rate in a cue count analysis, tick the "Cue rate" box

Note! You can only add multipliers if you have already created the appropriate fields in the Data Explorer.

Tip! If you used the Setup Project Wizard to define your multiplier fields, then they will appear automatically in the Multiplier tab in Model Definition Properties. For these fields, Distance also knows whether the operator is * or / (i.e. whether to multiply or divide the density estimate).

Variance Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

In the Variance page you specify the methods of calculating the variance of the density estimate. For the analytic variance estimate, you can choose the method for calculating the encounter rate variance component. You can also tell Distance to calculate a bootstrap variance estimate, and specify the exact methods to use.

Analytic Variance Estimate

Note! The analytic variance estimate is usually calculated automatically. It is not calculated in MCDS analyses when cluster size is a covariate or when the detection function is fit at one level and estimated at a lower level.

The encounter rate variance can be calculated in three ways (see Buckland et al. 2001 section 3.6.2, and look in the book index under Poisson variance of n)

- Estimate variance empirically. This is usually the best option the variance is calculated from the variance in observations between samples. However this is unreliable when there are few samples.
- Assume distribution of observations is Poisson.
- Assume distribution is Poisson, with overdispersion factor b. Setting b to 1 is equivalent to the previous option..

Bootstrap Variance Estimate

Check on the box **Select non-parametric bootstrap** to tell Distance to estimate the variance from bootstrap resamples of the data. When you run an analysis with this option checked in the Model Definition, the bootstrap results are given at the end of the Results tab on the Analysis Details. Distance reports bootstrap confidence limits using two methods: firstly using the bootstrap estimate of variance but assuming that the distribution of the density estimate is log-normal; secondly using the bootstrap percentile method (i.e., gives the appropriate quantiles of the actual bootstrap estimates). Distance also reports the mean of the bootstrap point estimates.

Each bootstrap resample is made up by sampling with replacement an equal number of units at the level you specify. For example, if you specify to resample samples (see below), then each bootstrap resample is made up of the same number of samples (line or point transects) as your original sample, chosen randomly with replacement from the original sample. Note that for line transects, this means that the survey effort (total line length) will differ in each resample. Note also that each of your original samples has an equal probability of appearing in the resample (an alternative, which we do not implement, would be to have probability proportional to line length).

Tip! You can add a column for the bootstrap CV and confidence limits in the Analysis Browser using the Analysis Browser Column Manager.

The options under Levels of Resampling change, depending on whether you have stratification turned on or not in the Estimate tab.

With **No stratification** you can resample samples and/or resample observations within samples. In this case you will normally want to select Resample samples alone. The resampling observations option is included for completeness but its routine use is not recommended, and can only be expected to produce reasonable results if the number of observations per sample is reasonable (e.g.., > 15).

Tip! If your survey includes stratification, but you want to resample across strata then choose **No stratification** in the Estimate tab, and then choose **Resample samples** in the Bootstraps **Levels of sampling** box.

With **Stratification or Post-stratification** enabled in the Estimate tab, you can resample by strata, samples or observations within samples. Normally, you will want to resample samples within strata. Resampling strata in addition could be useful if density is estimated by stratum or if sampling was stratified a priori.

Warning! If you are estimating variance using the bootstrap, be aware that the variance due to any multipliers is *not* included in the bootstrap variance. For more on multipliers, see Multipliers in CDS Analysis in the Users Guide.

Bootstrap Statistics File

Distance can save a file of summary statistics for each bootstrap iteration. This can be useful if you are having problems with the bootstrapping option. Check the Create file of statistics option and choose the file name using the Browse button.

For more information about the format of this file, see Saving CDS results to file in Chapter 8 of the Users Guide.

One reason to use bootstrapping is for multi-model inference – see Model Averaging in CDS Analysis in Chapter 8 of the Users Guide.

Misc. Tab - CDS and MCDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

This page contains miscellaneous Model Definition options regarding the presentation of results, saving of output to files, and options for smearing distance data before analysis.

Presentation of Results

Use the confidence intervals option to select the percentile of the two sided confidence intervals presented in the Results section of the Analysis Details window.

The option to report results for each iteration of the detection function fitting routine gives you more information about the fitting process, in the Detection Function pages in the Results section of the Analysis Details window. Use this option when there are problems with the fitting algorithm and you want more information about what has gone wrong. It is checked by default for MCDS analyses.

Results Files

These options allow you to save the results to files when the analysis is run. Two files can be saved:

- The results details file is identical to the text that is produced in the Results tab of the Analysis Details window.
- The results stats file is a compact output of summary statistics that Distance uses internally to extract data for the Analysis Browser table.

These files may be useful if you wish to import the results into another package - for example you could write a spreadsheet macro to parse the files and extract information into spreadsheet cells. For more information about these files, and their format specification see the Users Guide page Saving CDS results to file in Chapter 8.

Smearing

Smearing is an ad-hoc method for dealing with measurement error in line transect surveys (see Buckland et al. 2001 p 269-271).

The smearing option in Distance is only enabled if the survey is a line transect with distances collected as radial distance and angle. In addition, you should only select this option of you are going to pair the Model Definition with a Data Filter in which you have selected automatic Intervals. (In the **Intervals** tab of

the **Data Filter**, tick **Transform distance data into intervals for analysis**, and also **select Automatic equal intervals**.)

The smearing angle (phi) is the angle sector around the angle measurement. The proportion of distance parameter (s) is the proportional sector of distance to use as the basis for smearing. If an observation is measured at angle a and radial distance r, it is smeared uniformly in the sector defined by the angle range (a-phi, a+phi) and distance range (r*(1-s), r*(1+s)). The proportion of the sector contained in each perpendicular distance interval is summed as an observation frequency and these non-integer frequencies (grouped data) are analyzed to estimate detection probability.

Smearing is not supported in the MCDS engine

Model Definition Properties - MRDS

Estimate Tab - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

On the Estimate page you define the stratum and sample layer to use in the analysis, and tell Distance which quantities to estimate.

Stratum definition

Here, you specify the level of stratification to use in the analysis. For details, see Stratification and Post-stratification in MRDS in Chapter 10 of the Users Guide.

You specify layer types rather than layer names at this stage because it is only when the analysis is run that the survey object is used to select the data layers for the analysis. If you select a field for post-stratification that is not in the layers used in the analysis, an error will result.

Sample definition

Here, you specify which sample or sub-sample layer to use as the sample, for determining encounter rate variation - see Sample Definition in MRDS Analysis in Chapter 10 of the Users Guide for details.

Quantities to estimate

These options determine which quantities to estimate, and with what data.

The first option determines whether the engine should **Estimate density / abundance** or not. By default, this option is checked, but there are two circumstances under which you might want to uncheck it:

- In exploratory analysis, you might want to focus on fitting detection functions and leave the estimation of density until you've selected the detection function to use. This also saves computer time, since estimating density can be time consuming for larger datasets.
- You may wish to use different subsets of the data for estimating detection function and for estimating density given a detection – see below.

The second set of options determines how to obtain the **Detection function** parameters:

- The default option (**Estimate detection function**) is to fit the detection function using the data in this analysis. This makes sense most of the time.
- The second option is to **Use fitted detection function from previous MRDS analysis**. If you select this option, you must select the ID of the analysis containing the detection function you wish to use. For more about this option, and when you might want to use it, see Using a Previously Fitted Detection Function to Estimate Density in MRDS in Chapter 10 of the Users Guide (which needs finishing).

For the second option to work, you must have run the analysis containing the target detection function after un-checking the option in **Tools | Preferences | Analysis | R Software** to **Remove the new objects that are created with each run**.

When you choose the option to use the fitted detection function from a previous analysis, you cannot choose any options under the **Detection function** tab in the current Model Definition – since you are not fitting a detection function in this analysis.

Detection Function Tab - MRDS

The Detection Function tab is divided into five pages:

- **Method** - Detection Function Tab - MRDS
- **DS Model** - Detection Function Tab - MRDS
- **MR Model** - Detection Function Tab - MRDS
- **Factors** - Detection Function Tab - MRDS
- **Control** - Detection Function Tab - MRDS
- **Diagnostics** - Detection Function Tab - MRDS

Method - Detection Function Tab - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

On this page, you specify the **Fitting method** – see Introduction to MRDS Models in Chapter 10 of the Users Guide for more information about the methods available. Note that the choice of fitting method affects the options available in the subsequent pages, particularly the DS and MR model pages.

DS Model - Detection Function Tab - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

The DS model is the probability of one or more observer detecting the object, given it's distance and covariate values. On this page you specify the form of this model - the key function and any covariates that affect the scale parameter. For more about the form of the DS model, see Defining MRDS Models in Chapter 10 of the Users Guide. That section also includes a detailed description of how to specify the formula. Note that if you want covariates in the formula to be factor covariates, you need to specify them as factors – see Factors - Detection Function Tab - MRDS.

MR Model - Detection Function Tab - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

The MR model is the conditional detection function $p_{j|3-j}(y,\underline{z})$ – the probability of observer *j* detecting the object, given that the other observer (observer 3-*j*) has detected it and also given its distance and covariate values.. On this page you specify the form of this model. You specify the **Class of model** (currently only GLM), the **link function** (currently only logit), and the **formula** for the linear (or additive for GAM) predictor. For more about the form of the MR model, see Defining MRDS Models in Chapter 10 of the Users Guide. That section also includes a detailed description of how to specify the formula. Note that if you want covariates in the formula to be factor covariates, you need to specify them as factors – see Factors - Detection Function Tab - MRDS.

Factors - Detection Function Tab - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

On this page, you list the covariates in the DS and MR models that should be considered as factor covariates. Use a comma-delimited list, and use the same covariate names that appear in the DS and MR model formulae. For more on factor covariates, see Factor and Non-factor Covariates in MRDS in Chapter 10 of the Users Guide.

Control - Detection Function Tab - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

In this page, you can set various options to control the way the maximization routine performs in fitting the detection function models. For more details, see the section on Fine-tuning an MRDS Analysis in Chapter 10 of the Users Guide.

Diagnostics - Detection Function Tab - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

This page is the last in the MRDS Detection function tab. Here you can choose whether to plot the detection function histograms and perform the goodness of fit tests. These are done by default, but you may want to un-check the options to save time in an analysis, or save disk space in the case of the plots.

We anticipate that further options for setting Chi-square GOF cutpoints will be available in future versions of this engine.

Variance - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

In the Variance page you specify the methods of calculating the variance of the density and abundance estimates. The options are:

- Innes *et al.* (2002) – Based on the empirical variance of estimated density between samples (the default and preferred option).
- Buckland *et al.* (2001) – Based on the delta method, using the empirical variance in encounter rate between samples
- Binomial variance of detection process – Only realistic if the entire study area was sampled.

Misc - MRDS

See Model Definition Properties Dialog in the Program Reference for an overview of the Model Definition Properties dialog.

In the Misc page you specify various miscellaneous options.

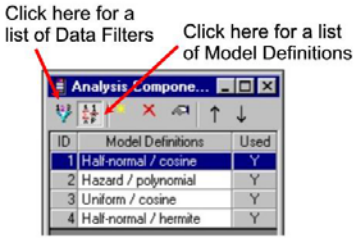
Presentation of results from density estimation

- **Standard output**. Gives default output options.
- **Extended output.** Gives extra output – see MRDS Results Details Listing in Chapter 10 of the Users Guide for details.

Model Definition Properties – DSM

Analysis Components Window

The Analysis Components window is designed to be a convenient way of manipulating Data Filters and Model Definitions. The window shows a list of all Data Filters or Model Definitions in your project. Using buttons on the toolbar you can create, delete, view and arrange the listed components.


Click here for a list of Data Filters
Click here for a list of Model Definitions
Analysis Compone...
ID
Model Definitions
Used
1 Half-normal / cosine Y
2 Hazard / polynomial Y
3 Uniform / cosine Y
4 Half-normal / hermite Y

Analysis Components window, showing a list of the Model Definitions in the Ducknest sample project

Tip! The last column of the in the table of analysis contents tells you whether that component is currently being used in any analyses: "Y" means it is being used and "N" means that it is not. This is useful because when there are many components (e.g., many Model Definitions if you have been doing a lot of analyses), it is easy to loose track of which are being used and which are no longer required. Also, if you double-click on a "Y", you get a list of the analyses that use that component.

Toolbar

- **List Data Filters**. When this button is selected, the Analysis Components window shows a list of the all Data Filters in the project.
- **List Model Definitions**. When this button is selected, the Analysis Components window shows a list of all the Model Definitions in the project.
- **New Item**. Create a new Data Filter or Model Definition, based on the one currently selected
- **Delete Item**. Delete the selected Data Filter or Model Definition

- **View Item Properties**. Show properties dialog for the selected Data Filter or Model Definition
- **Up**. Move the selected Data Filter or Model Definition up one place. This also
- **Down**. Move the selected Data Filter or Model Definition down one place.

For more information, see Analysis Components, in Chapter 7 of the Users Guide.

Other Windows

About Distance Dialog

This window reports information about the program such as the version number, program sponsors, authors and program files. The information is divided into a number of tabs:

- **About**. This shows the program version and release number. Use this when citing the program (click on the **Citation…** button for a suggested wording for the citation).
- **Sponsors**. Lists the program sponsors.
- **Authors**. Lists the program authors.
- **Use Agreement**. Contains a copy of the program use agreement. Acceptance of this agreement is a condition of program use.
- **Program Files**. Lists the files in the Distance program folder and gives information such as the version number of each file.

In addition, there are the following buttons at the bottom of the window:

- **Citation…** Click on this to obtain a suggested wording for citing Distance.
- **System Info…** Click on this to open a system tool that allows you to retrieve detailed information about the computer running Distance.
- **OK**. Click on this to close the About Distance dialog.

Export Project Dialog

The export project dialog allows you to export projects to another location on your computer. For more information about why you would want to do this, and some useful tips, see the Users Guide section on Exporting, Transporting and Archiving Projects (Chapter 4).

Options

File name: Enter the name of the project or zip file to save to

Save as type: You can export either as a Distance project (i.e., .dst project file and associated .dat data folder) or as a single .zip archive file.

Exclude the following parts: By default, the entire contents of the project are exported, but you can optionally exclude the following parts:

- **Data layer contents.** Checking this option causes the data to be excluded from the exported project, but the data structure (i.e., the data layer names and types, the data fields, etc) is retained. You

typically do this to save space if you are exporting a project to make a template for future project setup – see the Users Guide (Chapter 4) on Using an Existing Project as a Template.

- **Design and survey results.** Checking this option causes all designs and surveys in the project to be reset – i.e., their status is set to Not Run and any results are deleted. However, the specifications (the stuff on the Inputs tab of the Details page) are retained. You typically do this if you are making a template, or if you want to save space when transporting a file, and the results will be easy to re-create.
- **Analysis results.** Checking this option causes all results in the project to be reset – i.e., their status is set to Not Run and any results are deleted. However, the specifications – including all Data Filters and Model Definitions – are left untouched. You typically do this for the same reasons as you exclude the design and survey results.

If you are exporting to a zip archive on a removable disk (e.g., a floppy disk), and the archive is too big to fit on one disk, Distance will automatically span the archive across multiple disks. You will be prompted to insert one disk at a time, until the operation is complete. When opening a spanned archive from disk, you should insert the last disk first.

Projection Parameters Dialog

The Projection Parameters dialog is displayed by clicking the **Parameters…** button beside the projection lists in the Preferences or Project Properties dialogs. It allows you to set parameters associated with a particular projection. Projection parameters go with a projection and geographic coordinate system to make a projected coordinate system.

To set a parameter, tick the Set column and enter the appropriate value in the **Value** column.

Create New Layer Dialog

The Create New Layer dialog is accessed from the Data Explorer by clicking on the Create New Data Layer button [picture here], or from the menu **Data** | **Create Data Layer…**.

You are prompted to enter the **Layer name**, **Parent layer name**, and **Layer type**. The list of Layer types depends on the Parent data layer – for example a stratum data layer must have a global layer as parent.

For layers of type Coverage, the **Properties…** button allows you to access the Grid Properties dialog.

Grid Properties Dialog

The Grid Properties dialog is accessed by clicking the **Properties…** button of the Create New Layer dialog, when you are creating a new coverage data layer.

The settings you choose here are used to generate a grid of points used to assess probability of coverage for survey designs. For more information, see Concept: Coverage Probability in Chapter 6 of the Users Guide.

Projection for grid calculations. This is set to the default projection (see Project Properties), or [None] if the coverage layer doesn't have a coordinate

system. For more about coordinate systems and projections, see Coordinate Systems and Projections in Chapter 5 of the Users Guide.

Distance between grid points. Set this to the distance you want between grid points in the coverage layer. For more about selecting an appropriate distance, see Concept: Coverage Probability in the Design Properties section of the Program Reference.

Units of distance. If the calculations are to be done projected, or there is a non-earth geographic coordinate system, then you can choose from linear units (e.g., meters, miles, etc). Otherwise, they are angular (e.g., degrees of latitude or longitude).

Insert or Append Field Dialog

This dialog allows you to insert or append a field into the Distance database. It is accessed by clicking on the **Insert New Field Before Current** or **Append New Field After Current** buttons or menu items in the Data Explorer.

For more information about creating fields, see Data Explorer – Editing, Adding and Deleting Fields in the Program Reference.

Options

Field name: Enter the name of the new field here

Field type: Choose an appropriate field type for the field – note that this cannot be changed once the field has been created.

Units: Choose units for the new field, if appropriate. The units can be changed later in the Data Explorer (see Data Explorer – Editing, Adding and Deleting Fields in the Program Reference).

Data Layer Properties Dialog

This dialog gives you information about a particular data layer. It is accessed by selecting a data layer in the Data Explorer and clicking on the button, or choosing the menu item **Data** | **Data Layer Properties…**

There are two tabs:

- **General** – gives information about the **Layer name**, **Layer type** and **Parent layer**. The **Description** box allows you to enter comments about the data layer. For coverage layers, this box contains information about the options used to generate the coverage.
- **Geographic data** – gives information about the geographic information associated with the layer, if any. For more information about how GIS data are stored, see Geographic (GIS) Data in Chapter 7 of the Users Guide.
  - **Shapefile** is the name of the ESRI shapefile containing the geographic data.
  - **Shape type** gives the type of shape (point, line or polygon)
  - **Folder** gives the Windows folder containing the shapefile – by default this is the project's Data Folder – but shapefiles can be contained in any folder (see Importing Existing GIS Data in Chapter 5 of the Users Guide)
  - **Coordinate system** gives details of the coordinate system of the shapefile. Click on **Change coordinate system …** to go to the New Coordinate System dialog, where you can specify a different coordinate system for the shapefile data.

Shape Properties Dialog

The shape properties dialog allows you to view and edit the vertices (corners) of a shape attached to a record in a data layer, as well as copy the vertices to the Windows clipboard and paste them from the clipboard. It is accessed from the Data Explorer, by double-clicking on a record in the "Shape" field.

There are three types of shape:

- Point – has only one vertex
- Line – a series of points, joined up. A multi-part line is a broken line, e.g.: —— ——. The break is introduced using a Separator in the list of vertices.
- Polygon – a solid shape, made up of a set of vertices. Multi-part polygons can be created using separators in the list of vertices.

The dialog contains a table listing the x and y coordinates of the vertices making up the selected shape. Coordinates are in units defined by the coordinate system of the data layer (for more about coordinate systems, see Coordinate Systems and Projections in Chapter 5 of the Users Guide).

You can edit the coordinates by typing into the table entries. You can add vertices by clicking on the **Insert vertex** and **Append vertex** buttons. You can delete a vertex by highlighting it and clicking on the **Delete** button. You can delete all the vertices by clicking the **Delete All Vertices** button.

To create multi-part lines or polygons, highlight the last vertex in the line or polygon and click the **Append Separator** button. Then click **Append point** to create the first point of the new line or polygon.

You can copy the vertices to the Windows clipboard by clicking **Copy to Clipboard**, and then paste this into another file such as a text file or spreadsheet. Each vertex is copied to the clipboard as two numbers separated by a tab character. Separators between multi-part lines or polygons are copied a line containing just a tab character.

You can also paste the vertices of a shape from the Windows clipboard into the table by clicking **Paste from Clipboard**. This provides a useful mechanism for importing shape information from other packages such as text files or spreadsheets. For more details, see Importing GIS Data via the Windows clipboard in Chapter 5 of the Users Guide.

Once you have finished editing the point, click **OK** to save the edits and return to the Data Explorer. Alternatively, click **Cancel** to cancel any changes and return.

Take care when editing shapes, as once you have pressed **OK** there is no undo button!

New Coordinate System Dialog

This dialog lets you change the coordinate system of a data layer. It is accessed from the **Geographic** tab of the **Data Layer Properties** dialog, by clicking on **Change coordinate system…**.

For more information about coordinate systems in Distance, see Coordinate Systems and Projections in Chapter 5 of the Users Guide)

Column Manager Dialog

The Column Manager allows you to add, delete and rearrange the summary columns of results that appear in the Design, Survey and Analysis Browser. Each set can have different results columns. For example, in analysis, you could have a set for exploratory data analyses, with columns such as number of parameter, Delta AIC and Chi-sq *p*, and another set for final results, with columns such as N and CV of N. (For more information about some of the columns in the Analysis Browser, see CDS Analysis Browser Results in Chapter 8 of the Users Guide).

The Column Manager consists of two tables - the left table lists the columns that are already included and the right table lists those that are available to be included. Each column shows the column name, as it will appear in the browser, and an explanation of the meaning of the column.

To include a column, click on the column name in the available table and press the **<** button. Double clicking on the column name has the same effect.

To exclude a column, click on the column name in the selected table and press the **>** buttons. Double clicking on the column name has the same effect.

You can select more than one analysis at once in either table, by holding the **Ctrl** or **Shift** keys while you click, or by pressing **Ctrl A** or **Ctrl /** to select all the analyses.

To rearrange the ordering of the columns in the selected table, use the ↑ and ↓ buttons.

To reset the columns to their state when the Column Manager was opened, press the **Reset** button.

To reset to columns to their default arrangement, press the **Default** button. You can edit the default arrangement in the Preferences dialog (choose **File** | **Preferences…** on the main menu).

To leave the Column Manager without saving the changes, press the **Cancel** button.

To save the changes and exit the Column manager, press **OK**.

Arrange Sets Dialog

This dialog lets you change the order that Design, Survey and Analysis sets appear in the drop-down list in the Design, Survey and Analysis browsers. You access the dialog by clicking on the **Arrange** Set**s** button on the browser toolbar.

For more about sets, look under the appropriate browser (Design, Survey or Analysis) in the Project Browser section of the Program Reference.

Map Properties Dialog

The Map Properties dialog displays information about a map. It is accessed by clicking the **Map Properties** button on the Map toolbar. Information displayed includes the map's background colour, rotation angle and coordinate system.

Add Map Layer Dialog

The Add Map Layer dialog allows you to choose which layer to add to the map from the drop-down list. It is accessed by clicking on the **Add Layer** button in a Map.

Run Design Dialog

The Run Design dialog opens when you run a design. Here, you choose whether to estimate probability of coverage for your design, or create a new survey based on the design.

Confirm Change Dialog

This dialog is displayed:

- when you change the properties of a design that is being used by one or more survey objects that have been generated from the design
- when you change the properties of a survey, data filter or model definition that is being used by an analysis that has been run and has results.

In it displays a list of surveys or analyses that will be affected by the change, and asks you to confirm that the log and results for these should be deleted and their status reset to "not run".

For more information about why the status needs to be reset, see Analysis Components in Chapter 7 of the Users Guide.

R Image Properties Dialog

This dialog is displayed by clicking the **Image Properties…** button in the **Analysis** tab of the Preferences Dialog. It allows you to change the properties of the images produced by the MRDS analysis engine. If you are familiar with the package R, you will recognize many of these options as relating to the `par` function in R.

Image file format

Images are saved to files in a folder "R" within the project data folder - for more on this see Images Produced by R in the Users Guide. Use these options to alter the format of this file. Some possible reasons for changing the options are:

- you want to use the image files for inclusion in another document and would like them to be a specific format or size
- you want to make your project file smaller, by specifying smaller images, or a format where the images are compressed (jpg)
- your operating system is having trouble displaying images in the Distance interface in the default format (wmf), so you want to switch to another

The options are:

- **Format of image file** – three formats are currently available:
  - **wmf** – Windows metafile is a vector-based format, and gives the best display quality in the Distance interface, where the graphic gets resized according to the size of the Analysis Details window. It is therefore the default. Wmf files are, however, relatively large (compared with jpeg). There may also be problems displaying wmf files on older operating systems (e.g., Windows 98, Windows NT), in which case one of the other formats will have to be used.
  - **jpeg** – A compressed pixel-based format. Produces relatively small images, but they tend to look poor when rescaled.

        - **bmp** – A non-compressed pixel-based format. Produces large images that tend to look poor when rescaled, but is a format that can be viewed in any operating system.
    - **Character point size** – default point size of plotted text, interpreted at 72 dpi, so one point is approximately one pixel
    - **wmf**
        - **width** – the width of the plot in inches
        - **height** – the height of the plot in inches
    - **jpeg**
        - **width** – the width of the plot in pixels for jpeg and bmp formats
        - **height** – the height of the plot in pixels for jpeg and bmp formats
        - **quality** – the the "quality" of the jpeg image, as a percentage. Smaller values will give more compression but also more degradation of the image.

Graphics parameters

These options allow you to change the look of the plot. They correspond with options available in the par statement in R. More could be added in future – just ask!

- **Line width** (lwd) – The line width, a positive number, defaulting to 2. We use a default of 2 as it looks better in the wmf plots.
- **Line type** (lty) – The type of line to draw. An integer: 0=blank, 1=solid (the default), 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash.
- **Point character** (pch) – An integer specifying the symbol to use – try some numbers between 1 and 20 to see what it does. The default is 1, a filled circle.
- **Symbol scaling** (cex) – A numerical value giving the amount by which plotting text and symbols should be scaled relative to the default. Default is 1.

Data Selection Zoom Dialog

This dialog is displayed by pressing SHIFT-F1 (i.e., the shift and F1 keys simultaneously) while editing a data selection criterion in the Data Selection Tab of the Data Filter Properties Dialog. Here, you have more space to view and edit long, complex selection criteria.

For more about the format of data selection queries, see the page on the Data Selection Tab in the Program Reference.

Appendix - Inside Distance

Introduction to Inside Distance Appendix

Advanced Topic

This chapter contains information about how Distance works from the inside. It is intended for advanced users who want to push Distance to its limits.

The information here is preliminary, and will be expanded in future releases.

Warning! Manually editing the distance database files can result in the project no longer working in Distance. The following information is provided in the hope that it will be useful, but we cannot be held responsible for any problems that occur as a result of using it.

Distance components

Advanced Topic

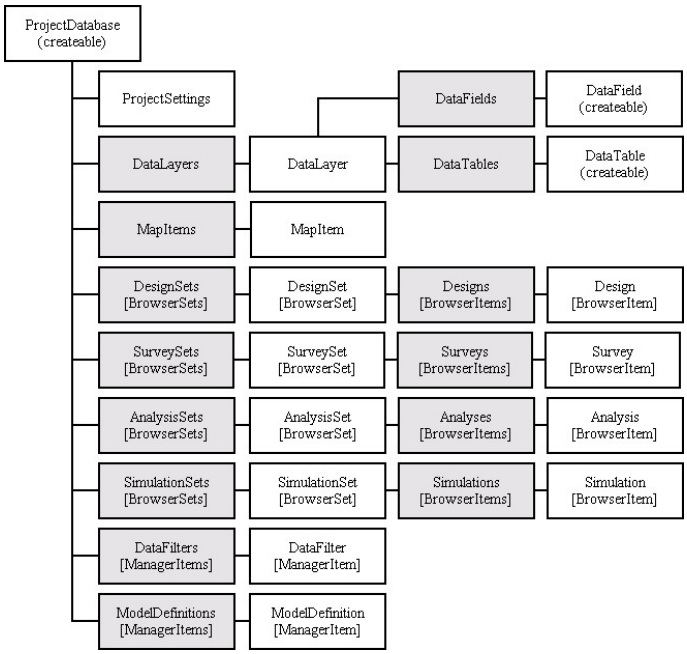
This section will contain a description of the various components that go to make up Distance. You can get a list of the component files that make up Distance by selecting the Help | About Distance… from the main menu, and clicking on the Program Files tab.

The Distance 6 Database API

Advanced Topic

All of the functionality used to manipulate Distance 6 projects is packaged into an ActiveX DLL, D6DbEng.dll. This DLL can be used to create and delete projects, change project and default properties, create, delete and rename data layers, etc. An overview class diagram for the API (Application Programming Interface) is given below, and detailed documentation is available from the program authors on request.

Warning! The database API is primarily intended for internal use by those working on the Distance project. There is no guarantee that the API will remain the same in subsequent releases, although we will endeavor not to break the current interface.



Overview of D4DbEng.dll public classes. Where there are items in [], these are the base class names.

Data File Reference

Advanced Topic

How Distance Stores Data

Distance can use data from three sources:

- internal data, which is stored in tables in the Data File, DistData.mdb
- geographic data, stored in ESRI shapefiles
- external data, stored in external database tables, spreadsheet or text files

Information about the location of the data is stored in three linked tables in DistData.mdb. This file is created by the Microsoft Jet 3.51 engine, which is the same engine used in Microsoft Access 97. If you want to get at the internal workings of this file, your best bet is to use Access 97. You can use Access 2000 or later, but things are a little more complex (see Accessing DistData.mdb using newer versions of Access.)

The three tables in DistData.mdb are:

- **DataLayers** – contains one record for each data layer in the project
- **DataTables** – contains one record for each database table that make up the data layers. For example, by default a geographic data layer is made up of two tables: an internal table with the same name as the data layer, and the .dbf table from the shapefile.
- **DataFields** – contains one record for each data field in the data layers of the project.

More details on each of these tables are given in the following topics.

DataLayers table in DistData.mdb

This table contains one record for each data layer in the project. It is used to tell Distance how many layers are in the project and what their relationship is to one another - this is used, for example, to construct the tree view in the Data Layers Viewer.

The **DataLayers** table has the following fields:

Field Name	Field Type	Primary Key?	Description
LayerName	Text	Y	Name of data layer
LayerType	Long	N	Enumeration (see Enumerations in DistData.mdb)
ParentLayerName	Text	N	Name of parent data layer
Description	Memo	N	Place for user to enter comments about the layer (can edit in Description box of the Data Layer Properties Dialog).

Records in the DataLayers table are subject to the following restrictions:

- The LayerName must be unique
- There are some restrictions on characters that can be used for layer names (e.g., no special characters such as "[", "]", etc.
- There can only be one record with LayerType = 1 (Global)
- All layers, except the global layer, must have a ParentLayerName.
- In general, the layer type of the parent layer should be lower than that of its children (e.g., parent LayerType = 1 (Global); child LayerType = 2 (Stratum)).

DataTables table in DistData.mdb

This table contains one record for each database table that makes up each data layer in the project. For example, by default a geographic data layer is made up of two tables: an internal table with the same name as the data layer, and the .dbf table from the shapefile. Fields from tables making up the data layer are joined together to make up the records for the data layer that you see in the Data Explorer.

The **DataTables** table has the following fields:

Field Name	Field Type	Primary Key?	Description
TableName	Text	Y	Unique name for table. Can be either an internal table, or a linked table. This name is assigned by Distance when the table is added to the project – if you're adding tables manually you can make a unique name up.
LayerName	Text	N	Name of data layer
SourceDatabaseType	Text	N	Either (1) "Int" for internal tables (i.e., tables that are in DistData.mdb); (2) "Geog" for the .dbf part of a shapefiles; or (3) a MS Jet IISAM database type specifier (e.g., "Text")

			(see Note 1).
SourceDatabaseName	Text	N	Either (1) blank if SourceDatabaseType is "Int"; (2) blank if the table is in the Data Folder; or (3) the absolute path to the database (see Note 2).
SourceTableName	Text	N	The actual table name (see Note 3).
ShapeType	Long	N	Enumeration (see Enumerations in DistData.mdb)
PrimaryTable	Boolean	N	If true, the table should contain ID and ParentID fields. If false, the table should contain LinkID field. Only one primary table per data layer.

Note 1: MS Jet IISAM specification is for this to be blank for native Jet databases, but this property should contain "Jet" if the table is a native Jet table.

Note 2: for single table databases, this is the folder name, for multiple table databases, its the file name

Note 3: For external files: for single table databases, this is the file name without extension, for multiple its the table name in the file.

Records in the DataTables table are subject to the following rules:

- The TableName must be unique, and there are some restrictions on the characters that can be used.
- There must be at least one table with SourceDatabaseType="Int" (i.e., internal) in each data layer
- One table in each layer must be the primary table (PrimaryTable=T) and this table must be of SourceDatabaseType="Int".
- There can be at most one geographic table (SourceDatabaseType="Geog"). The geographic table must have a nonzero ShapeType.

Linking to external tables using MS Jet IISAM is examined further in Linking to External Data from DistData.mdb.

DataFields table in DistData.mdb

This table contains one record for each field in each data layer. It is used to determine which fields to display in the Data Explorer, and also to denote special fields such as ID fields.

The **DataFields** table has the following fields:

Field Name	Field Type	Primary Key?	Description
LayerName	Text	Y	Name of data layer
FieldName	Text	Y	Name of data field (must be the same as the name of the field in the source database table)
TableName	Text	N	TableName in DataTable

FieldType	Long	N	Enumeration (see Enumerations in DistData.mdb)
Units	Text	N	Units of measurement[1], or blank.
OrdinalPosition	Long	N	Used in Data Explorer to tell what order to display the fields
ColumnWidth	Long	N	Used in Data Explorer to remember user-specified column widths (usually zero)
Tag	Text	N	Currently unused
Formula	Memo	N	Currently unused

1 For a list of allowable strings, open DistIni.mdb in the Distance program directory, look in the ProjectSettings table under section UnitsAbbr. All the values under Key are allowable units.

Records in the DataFields table should abide by these rules:

- Records can have the same field name (e.g., "ID"), but the same field name cannot be used twice within a layer (i.e., you can't have the same field in two Data Tables that make up the same layer)
- Not all fields that physically exist in a Data Table need to have records here – if a field is omitted then it will not be available to Distance (e.g., it will not appear in the data sheet)
- In general, the FieldType must match the type of the field in the database (e.g., you'll get an error if a field that is a text field in the database is given FieldType = 1 (Integer)).
- Primary tables must have one field of FieldType = 10 (ID). All other tables must have one field of FieldType = 13 (LinkID). These fields are used to join together records from each table. In general, the ID and LinkID values in each table should go from 1 to the number of records in the table.
- One of the tables in each layer must contain a field of FieldType = 11 (ParentID). This field is not needed in each table in a layer – only in one of them. The only exception is tables in the global layer: none of these need a ParentID field (since there is no parent layer). The ParentID field tells Distance which record in the parent data layer to put the child record in.
- If there is a geographic table in the layer, there needs to be a record with FieldName = "Shape", and FieldType = 14 (Shape).
- The OrdinalPosition should be unique for fields within a layer – these dictate the order that the fields appear in the Data Explorer. ParentID and LinkID fields are not shown in the DataExplorer, so their ordinal position doesn't matter.

Unfortunately, at present, you cannot include attribute fields from the geographic table – the Shape field is the only one you can include. This is due to a bug in the GIS component from ESRI and we hope to resolve this at some point in the future. Until then, this means there is no way to link to attribute data in the shapefiiles.

Enumerations in DistData.mdb

Field Name	Enumeration	Meaning
LayerType	1	Global
	10	Stratum
	11	SubStratum1
	12	SubStratum2
	13	SubStratum3
	14	SubStratum4
	15	SubStratum5
	20	Sample
	21	SubSample1
	22	SubSample2
	23	SubSample3
	24	SubSample4
	25	SubSample5
	30	Observation
	40	Coverage
	99	Other
FieldType	1	Integer
	2	Decimal
	3	Text
	10	ID
	11	ParentID
	12	Label
	13	LinkID
	14	Shape
	99	Other (not used)
ShapeType	0	No shape
	21	Point
	22	Line
	23	Polygon
	24	Multipoint (not used)
	25	Rectangle (not used)
	26	Ellipse (not used)

Accessing DistData.mdb using newer versions of Access

The first time you open a DistData.mdb file in versions of Access after Access 97 (e.g., Access 2000, 2002, etc.), it asks you if you want to convert the file to the new format, or open as is.

If you choose to open as is, you will get a message saying that you cannot change the database structure. This means that you cannot add new fields or new tables to the database, but you can edit records or add new records. In many cases (for example, see Importing Existing GIS Data in Chapter 5 of the Users Guide) this is fine. In general we recommend this as the easier option.

If you choose to convert the database, you will be prompted for a new filename to save the converted database to (e.g., DistData.new.mdb). Once the conversion has taken place, the new file is opened. You are then will be free to make any changes you want to the database structure, such as adding new tables, fields, etc. However, you then need to convert the database back to the old format before it can be opened in Distance.

To do this:

- Rename the original DistData.mdb file (e.g., to DistData.bak.mdb), and keep it as a backup in case something goes wrong
- In Access 2000 or 2002, on the **Tools** menu, point to **Database Utilities**, click **Convert Database**, and then click **To Prior Access Database Version**.
- In the **Convert Database Into** dialog box, type DistData.mdb in the **File name** box, and then click **Save**
- Close the database in Access 2000/2002
- Open the project file in Distance, and check that the changes you made have registered correctly
- Assuming everything is fine, you can delete the Access 2000/2002 database (e.g., DistData.new.mdb) and the old backup file (DistData.bak.mdb)

As you can see, this is quite a hassle! There are other issues to consider in making the conversion – for more information, look in the Access 2000 online help under "Convert an Access 2000 database to Access 97" and "About converting an Access 2000 database to Access 97" (Access 2002 and later have similar topics).

Aside! Visual Basic 5 professional came with source code for a demonstration database access program called VisData, which allows access to Microsoft Jet 3.51 files. It isn't nearly as easy to use as Access 97, but is better than nothing. If it would be useful, we could probably post a modified version of the program on the Distance web site. We'd have to check the licensing situation first though. If you think this would be useful, please let us know.

Linking to External Data from Distdata.mdb

Warning! This is an advanced technique, suitable only for those comfortable poking around inside Access databases. Make sure you have read and understood the previous topics on How Distance Stores Data before proceeding!

You can, in theory, use Distance to link to data in tabular text files, databases and spreadsheets – although you cannot do this directly from the Distance GUI. Instead, you do it by directly editing the Data File, DistData.mdb, using Access. Briefly: you add an entry for the table you want to link in the DataTables table in DistData.mdb, and then add entries for the fields you want to link to in DataFields. An example is provided in the LinkingExample sample project.

Supported External Data Sources

You can link to any type of data for which there is a Microsoft Jet 3.51 IISAM (Installable Indexed Sequential Access Method) driver. By default, Distance supplies you with:

- a native Jet driver (for Microsoft Access 97 and earlier .mdb databases)
- a text driver (for text files in tabular formats)
- a Microsoft Excel driver (for versions 3.0-8.0).

For more on the last two, see the topic Working with the Microsoft Jet IISAM Text File Driver and Working with the Microsoft Jet IISAM Excel File Driver.

There are also drivers available for the following formats:

- Microsoft FoxPro databases, versions 2.0, 2.5, 2.6, 3.0, and DBC (database containers).
- dBASE databases, versions III®, dBASE IV®, and dBASE 5.0.
- Paradox databases, versions 3.*x*, 4.*x*, and 5.*x*.
- Lotus spreadsheets, versions WKS, WK1, WK3, and WK4.
- Tabular data in Hypertext Markup Language (HTML) files.

These are discussed briefly in Working with Other Microsoft IISAM drivers.

The technology used by the Distance database engine (Jet 3.51) has been replaced by Microsoft by newer technology, so it is unlikely they will issue IISAM drivers to link to newer versions of the above software. Given the overhead that would be required, it is also unlikely that we will be updating the Distance database engine to use newer technology any time soon. Many newer programs can, however, work with files in the older formats – for example, newer versions of Excel can easily save files as Excel 97-2002 (Excel 8.0) and work with them in that format.

Outline of Linking to External Data

Probably the easiest way to see how to link to external data is to examine the LinkedExample sample project. This project has two data layers:

- a global layer which links to a table in an Access 97 database (LinkedData.mdb)
- a stratum layer which links to another table in LinkedData.mdb and also to a tab-delimited text file, and to a worksheet in an Excel 8.0 file (LinkedData.xls).

Both layers also have a geographic data layer. Examine the DataTables and DataFields tables in DistData.mdb to see how this was done.

Imagine that we wanted to add a sample-level data layer called “Line transect” and link to a table “Transect” containing transect information in LinkedData.mdb. The following outlines how we might do this, from within DistData.mdb. For simplicity, we’ll assume that the new layer is not going to be geographic.

20. Create a new record in DataLayers for the new layer. LayerName is “Line transect”, LayerType is 20 (Sample) and ParentLayerName is “Region”.
21. Create a primary data table for the new layer. Create a new table called “Line transect” in DistData.mdb. Give it fields “ID” and “ParentID” (both of type Long). (Note – I’m assuming the ParentID field isn’t in the external “Transect” table.) Let’s imagine that there are 10 records in our external Transect table – so we need

10 records in this table, with ID from 1-10 and appropriate ParentID values to put the transects in the correct stratum.

22. Create a record for the primary data table in the DataTables table. The TableName, LayerName and SourceTableName are all “Line transect”, the SourceDatabaseType is “Int”, and PrimaryTable is True.

23. Create records for the fields in this table in DataFields. You need a record for the ID field, which will have FieldType=10 (ID) and for the ParentID field, which will have FieldType = 11 (ParentID)

Note that all of these steps could have been more easily performed by opening the project in Distance, creating a new data layer called “Line transect” of type Sample, and then creating 10 new records in the layer.

24. Now we want to add a new record to the DataTables table for the external table. The TableName can be anything, but for consistency with what’s already in the project let’s use “lnkTransect”, the LayerName is “Line transect”, SourceDatabaseType is “Jet”, SourceDatabaseName is “LinkedData” and SourceTableName is “Transect”.

25. Now we must add records to the DataFields table for each field we wish to link to in the Transect table. We must include as a minimum a field of type 13 (LinkID), which tells Distance which records in Transect link to which records in the internal “Line transect” table. We can include any other fields we like, so long as they are in the “Transect” table.

26. Hopefully now the new table will be linked – we can open the project in Distance to check.

If you run into problems linking files of a specific format, and have tried everything you can think of, try looking at the settings in HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Jet\3.5\Engines or \ISAM_Formats to see if they might be the cause of the problem.

Working With the Microsoft Jet IISAM Text File Driver

The following information is adapted from the Microsoft Jet 3.5 documentation, and will hopefully be of use in setting up text files for linking to the Distanace database. An example of linking to a text file is given by the LinkingExample sample project.

You can use the Microsoft Jet Text IISAM to link and open character-delimited and fixed-length text files. Commas, tabs, or user-defined delimiters are valid in the source file.

When specifying connection information for text files, use the following specifications in the DataTables table:

SourceDatabaseType: Text

SourceDatabaseName: The full path to the directory containing the text file you intend to access. If you do not specify a path, Distance will look in the project data folder.

SourceTableName: The name of the text file, including the extension. If you don't specify an extension, the default .txt extension is used.

Microsoft Jet recognizes null values in character-delimited files by the presence of two consecutive delimiting characters. Microsoft Jet recognizes null values in fixed-length files by the absence of data (spaces) in the data column.

Microsoft Jet determines the format of the text file by reading the file directly or by using a schema information file. The schema information file, which is always named Schema.ini and always kept in the same directory as the text data source, provides the IISAM with information about the general format of the file, the column name and data type information, and a number of other data characteristics. A Schema.ini file is always required for accessing fixed-length data; you should use a Schema.ini file when your text table contains DateTime, Currency, or Decimal data or any time you want more control over the handling of the data in the table.

Microsoft Jet doesn't support multiuser access to text files. When you open a text file through Microsoft Jet, you have exclusive access to the file.

Depending on the registry settings on your computer, you may only have read-only access to linked text files.

The following table lists the few limitations to the size of text tables and objects.

Item	Maximum size per text file
Field	255
Field name	64 characters
Field width	32,766 characters
Record size	65,000 bytes

Understanding Schema.ini files

Schema.ini files provide schema information about the records in one or more text files in the same directory as the schema file. Each Schema.ini entry specifies one of five characteristics of the table:

- The text file name.
- The file format.
- The field names, widths, and types.
- The character set.
- Special data type conversions.

The following sections discuss these characteristics.

Specifying the file name

The first entry in Schema.ini is always the name of the text source file enclosed in square brackets. The following example illustrates the entry for the file Sample.txt:

```
[Sample.txt]
```

You can specify settings for more than one file in the same Schema.ini file.

Specifying the file format

The Format option in Schema.ini specifies the format of the text file. The Text IISAM can read the format automatically from most character-delimited files. You can use any single character as a delimiter in the file except the double quotation mark ("). The Format setting in Schema.ini overrides the setting in the Windows Registry on a file-by-file basis. The following table lists the valid values for the Format option.

Format specifier	Table format
TabDelimited	Fields in the file are delimited by

	tabs.
CSVDelimited	Fields in the file are delimited by commas (comma-separated values).
Delimited(*)	Fields in the file are delimited by asterisks. You can substitute any character for the asterisk except the double quotation mark.
FixedLength	Fields in the file are of a fixed-length.

For example, to specify a comma-delimited format, you would add the following line to Schema.ini:

```
Format=CSVDelimited
```

Specifying the fields

You can specify field names in a character-delimited text file in two ways:

- Include the field names in the first row of the table and set ColNameHeader to True.
- Specify each column by number and designate the column name and data type.

You must specify each column by number and designate the column name, data type, and width for fixed-length files.

The ColNameHeader setting in Schema.ini overrides the FirstRowHasNames setting in the Windows Registry on a file-by-file basis.

You can also instruct Microsoft Jet to determine the data types of the fields. Use the MaxScanRows option to indicate how many rows Microsoft Jet should scan when determining the column types. If you set MaxScanRows to 0, Microsoft Jet scans the entire file. The MaxScanRows setting in Schema.ini overrides the setting in the Windows Registry on a file-by-file basis.

The following entry indicates that Microsoft Jet should use the data in the first row of the table to determine field names and should examine the entire file to determine the data types used:

```
ColNameHeader=True
MaxScanRows=0
```

The next entry designates fields in a table by using the column number (**Col***n*) option, which is optional for character-delimited files and required for fixed-length files. The example shows the Schema.ini entries for two fields, a 10-character CustomerNumber text field and a 30-character CustomerName text field:

```
Col1=CustomerNumber Text Width 10
Col2=CustomerName Text Width 30
```

The syntax of Col*n* is:

```
Coln=ColumnName type [Width #]
```

The following table describes each part of the Col*n* entry.

Parameter	Description
ColumnName	The text name of the column. If the column name contains embedded spaces, you must enclose it in double quotation marks.

Type	Data types are:Microsoft Jet data types: Bit, Byte, Short, Long, Currency,Single, Double, DateTime, Text, MemoODBC data types, Char (same as Text), Float (same as Double), Integer (same as Short), LongChar (same as Memo), Date date format
Width	The literal string value Width. Indicates that the following number designates the width of the column (optional for character-delimited files, required for fixed-length files).
#	The integer value that designates the width of the column (required if Width is specified).

Selecting a Character Set

You can select from two character sets: ANSI and OEM. The following example shows the Schema.ini entry for an OEM character set. The CharacterSet setting in Schema.ini overrides the setting in the Windows Registry on a file-by-file basis. The following example shows the Schema.ini entry that sets the character set to ANSI:

```
CharacterSet=ANSI
```

Specifying Data Type Formats and Conversions

The Schema.ini file contains a number of options that you can use to specify how data is converted or displayed when read by Microsoft Jet. The following table lists each of these options.

Option	Description
DateTimeFormat	Can be set to a format string indicating dates and times. You should specify this entry if all date/time fields in the import/export are handled with the same format. All of the Microsoft Jet formats except A.M. and P.M. are supported. In the absence of a format string, the Windows Control Panel short date picture and time options are used.
DecimalSymbol	Can be set to any single character that is used to separate the integer from the fractional part of a number.
NumberDigits	Indicates the number of decimal digits in the fractional portion of a number.
NumberLeadingZeros	Specifies whether a decimal value less than 1 and greater than –1 should contain leading zeros; this value can either be False (no leading zeros) or True.
CurrencySymbol	Indicates the currency symbol to be used for currency values in the text file. Examples include the

	dollar sign ($) and Dm.
CurrencyPosFormat	Can be set to any of the following values:· Currency symbol prefix with no separation ($1); Currency symbol suffix with no separation (1$); Currency symbol prefix with one character separation ($ 1); Currency symbol suffix with one character separation (1 $).
CurrencyDigits	Specifies the number of digits used for the fractional part of a currency amount.
CurrencyNegFormat	Can be one of the following values:· ($1);·–$1; $–1; $1–; (1$); –1$; 1–$; 1$–; –1 $; –$ 1; 1 $–; $ 1–; $ –1; 1– $; ($ 1); (1 $). This example shows the dollar sign, but you should replace it with the appropriate CurrencySymbol value in the actual program.
CurrencyThousandSymbol	Indicates the single-character symbol to be used for separating currency values in the text file by thousands.
CurrencyDecimalSymbol	Can be set to any single character that is used to separate the whole from the fractional part of a currency amount.

Note: If you omit an entry, the default value in the Windows Control Panel is used.

Examples of Schema.ini files

Schema.ini contains the specifics of a text data source: how the text file is formatted, how it's read at import time, and the default export format for files. The following example shows the layout for a fixed-width file, Filename.txt:

```
[Filename.txt]
ColNameHeader=False
Format=FixedLength
MaxScanRows=25
CharacterSet=OEM
Col1=columnname Char Width 24
Col2=columnname2 Date Width 9
Col3=columnname7 Float Width 10
Col4=columnname8 Integer Width 10
Col5=columnname9 LongChar Width 10
```

Similarly, the format for a delimited file is specified as:

```
[Delimit.txt]
ColNameHeader=True
Format=Delimited(!)
MaxScanRows=0
CharacterSet=OEM
Col1=username Text
Col2=dateofbirth DateTime
```

Note that both of these format sections can be in the same .ini file.

Another example of a Schema.ini file is in the Data Folder of the LinkingExample project.

Working With the Microsoft Jet IISAM Excel File Driver

The following information is adapted from the Microsoft Jet 3.5 documentation, and will hopefully be of use in setting up Excel spreadsheet files for linking to the Distance database. An example of linking to an Excel 8.0 file is given by the LinkingExample sample project.

The Microsoft Jet IISAMs support the following single-sheet worksheet and multiple-sheet workbook versions of Microsoft Excel: Excel 3.0 and Excel 4.0 for single-sheet worksheets, and Excel 5.0 (for Microsoft Excel 5.0 and 7.0) and Excel 8.0 for multiple-sheet workbooks. There are a few operations that you can't perform on Microsoft Excel worksheets or workbooks through the Microsoft Excel IISAM:

- You can't delete rows from Microsoft Excel worksheets or workbooks.
- You can clear data from individual cells in a worksheet, but you can't modify or clear cells that contain formulas.
- You can't create indexes on Microsoft Excel worksheets or workbooks.
- You can't read encrypted data through the Microsoft Excel IISAM. You can't use the PWD parameter in the connection string to open an encrypted worksheet or workbook, even if you supply the correct password. You must decrypt all Microsoft Excel worksheets or workbooks through the Microsoft Excel user interface if you plan to link or open them in your Microsoft Jet database.

When specifying connection information for Excel files, use the following specifications in the DataTables table:

SourceDatabaseType: One of: Excel 3.0, Excel 4.0, Excel 5.0, Excel 8.0

SourceDatabaseName: see Use (1) below

SourceTableName: see Use (2) below

To access this object	In this version of Microsoft Excel	Use this syntax
Entire sheet in a worksheet file	3.0 and 4.0	Use (1) to specify the fully qualified network or directory path to the worksheet file (no path needed if in data folder); use (2) to specify the sheet as *filename*#**xls**, where *filename* is the name of the worksheet.
Entire worksheet in a workbook file	5.0, 7.0, and 8.0	Use (1) to specify the fully qualified network or directory path to the workbook file (if not in data folder), including the workbook file name; use (2) to specify the sheet as *sheetname*$, where *sheetname* is the name of the worksheet. **Important** You must follow the worksheet name with a dollar sign ($).
Named range of cells in a worksheet or workbook file	3.0, 4.0, 5.0, 7.0, and 8.0	Use (1) to specify the fully qualified network or directory path to the worksheet or workbook file (if not in data folder), including the worksheet or workbook file name; use (2) to specify the named range as *NamedRange*, where

		NamedRange is the name you assigned to the range in Microsoft Excel.**Important** You must name the range in Microsoft Excel before attempting to open or link it.
Unnamed range of cells in a worksheet file	3.0 and 4.0	Use (1) to specify the fully qualified network or directory path to the worksheet file (if not in data folder), including the worksheet file name; use (2) to specify the range as **A1:Z256**. Replace **A1:Z256** with the range of cells you want to access.
Unnamed range of cells in a single worksheet in a workbook file	5.0 and 7.0	Use (1) to specify the fully qualified network or directory path to the workbook file (if not in data folder), including the workbook file name; use (2) to specify the sheet you want to link or open as *sheetname***$** and the range as **A1:Z256**. For example, to access cells A1 through Z256 in worksheet SheetName, you would use the following in (2) SheetName$A1:Z256.

The HDR parameter

By default, the first row of the worksheet or selected text contains the field name. To suppress this, add ;HDR=No to the SourceDatabaseType (you don't need to specify the default ;HDR=Yes, unless the default has been changed – this is governed by the value of FirstRowHasNames in the following registry key: \HKEY_LOCAL_MACHINE\SOFTWARE\ Microsoft\Jet\3.5\Engines\Excel)

Working with Other Microsoft IISAM drivers

Microsoft provides drivers for other data sources, in addition to text and native Jet 3.51 (see Supported External Data Sources for a list).

These drivers are not supplied with Distance. Probably the easiest way to obtain them is to go to the Support page of the Distance web site, and find the link to download Microsoft DAO setup package. As part of the install, it will prompt to ask which IISAM drivers you wish to install. You can also download them from Microsoft.

A very brief outline of what entries are required in the DataTables table is given below. More documentation can be supplied on request to the program authors; however it is not anticipated that this feature of the program will see much use!

Data Source	SourceDatabaseType	SourceDatabaseName	SourceTableName
Microsoft Jet	Jet	Path to the database	Name of the table
dBASE	One of: dBASE III dBASE IV dBASE 5.0	Path to the database	Name of the table
FoxPro	One of: FoxPro 2.0 FoxPro 2.5 FoxPro 2.6 FoxPro 3.0 FoxPro DBC	Path to the database (For Microsoft FoxPro DBC, the path must include the name of the .dbc file.)	The name of the table. (Use the .dbf file name without the extension, or the complete file name with the extension,

			but substitute a number sign [#] for the dot [.] that precedes the file name extension; for Microsoft FoxPro DBC, use the table name in the DBC.)
Paradox	One of: Paradox 3.x Paradox 4.x Paradox 5.x	Path to the database	Name of the table.
Lotus	One of: Lotus WKS[1] Lotus WK1[2] Lotus WK3[2] Lotus WK4[1]	Path to the file	filename#wks (single sheet file) or sheetname: (multi sheet file) or sheetname:A1..Z256 (range of cells)
HTML	HTML Import	URL of page containing table	title of table caption, or Table1, Table2, etc if no caption

[1] *Read-only access*
[2] *Read and insert access (can't modify existing rows)*

Valid Names

Valid Field Names

Field names must meet the following criteria to be valid:

- For internal fields, the name must be 64 letters long or less
- For shapefile fields, the name must be 10 letters or less long with no spaces
- The only permitted characters are letters (A-Z or a-z), numbers (0-9), spaces or underscores.
- Field names must be unique within a data layer (i.e., the same name is not allowed in 2 tables, except for the ID and LinkID fields)
- The name must not appear on the list of reserved field names below (not case sensitive).

Reserved field names		
Reserved by Distance		
Shape	FeatureID	None
Reserved by R		
NULL	NA	TRUE
FALSE	GLOBAL.ENV	Inf
NaN	function	while
if	repeat	for
in	else	next

break	...	
Reserved by the Jet database engine		
ADDALL	Alphanumeric	ALTER
And	ANY	AS
ASC	AUTOINCREMENT	Avg
Between	BINARY	BIT
BOOLEAN	BYBYTE	CHAR
CHARACTER	COLUMN	CONSTRAINT
Count	COUNTER	CREATE
CURRENCY	DATABASE	DATE
DATETIME	DELETE	DESC
DISALLOW	DISTINCT	DISTINCTROW
DOUBLE	DROP	Eqv
EXISTS	FLOAT	FLOAT8
FLOAT4	FOREIGN	FROM
GENERAL	GROUP	GUID
HAVING	IEEEDOUBLE	
IGNORE	Imp	In
ININDEX	INNER	INSERT
INT	INTEGER	INTEGER4
INTEGER1	INTEGER2	INTO
Is	JOIN	KEY
LEFT	Level	Like
LOGICAL	LOGICAL1	LONG
LONGBINARY	LONGTEXT	Max
MEMO	Min	Mod
MONEY	Not	NULL
NUMBER	NUMERIC	OLEOBJECT
ON	OPTION	Or
ORDER	Outer	OWNERACCESS
PARAMETERS	PERCENT	PIVOT
PRIMARY	PROCEDURE	REAL
REFERENCES	RIGHT	SELECT
SET	SHORT	SINGLE
SMALLINT	SOME	StDevStDevP
STRING	Sum	TABLE
TableID	TEXT	TIME
TIMETIMESTAMP	TOP	TRANSFORM
UNION	UNIQUE	UPDATE
VALUE	VALUES	Var
VARBINARY	VARCHAR	VarP
WHERE	WITH	Xor
YESNO		

Miscellaneous topics

Random number generation

Different algorithms are used to generate pseudo-random numbers in the different components of Distance, as follows:

- The Distance interface uses the random number class in D4Util.dll. This is based on Knuth's subtractive method – see algorithm `ran3` in Press et al. 1992. The generator is seeded from the system clock. Random numbers are used, for example, in generating probability of coverage grid points.
- The design engine also uses the random number class from D4Util.dll for generating designs. The generator is either seeded from the system clock or from a user-specified value.
- The CDS and MCDS analysis engines use the Compaq Visual Fortran function `random_number`, for bootstrap resampling. This uses two congruential generators (see L'Ecuyer 1988 or the Fortran manual for more details). The seed is either set from the system clock, or can be specified by the user in the **Model Definition Properties**, under **Variance**.
- The R routines in Distance use R's default random number generator – type **help(.Random.Seed)** in R for more information.

Appendix - MCDS Engine Reference

Introduction to MCDS Engine Reference

The CDS and MCDS engines are implemented as a stand-alone FORTRAN program, MCDS.exe. This program is called behind the scenes by Distance when you press the Run button on the Analysis Details Inputs tab. Some users may wish to run the engine from outside the Distance interface – either from the Windows command line or from another program. For example, you may want to automate the running of analyses for simulations, or you may want to run a complicated bootstrap not available in Distance. Here, we provide outline documentation for running the CDS and MCDS analysis engine as a stand-alone program.

For more information about the various options available in the engine, see the Users Guide Chapter 8 - Conventional Distance Sampling Analysis and Chapter 9 - Multiple Covariates Distance Sampling Analysis.

Since the CDS and MCDS analysis engines are both implemented in MCDS.exe, we refer to both as "the MCDS engine" in what follows.

There have been several messages on the distance-sampling email list providing tips on how to use the MCDS engine (and previous versions of the engine) from outside of Distance. Have a look in the online archives.

Some history

In historic versions of Distance (1.0 - 3.0), the program was driven by a simple command language, which defined the survey design, data, and analysis methods. Distance could be run in batch mode by passing in the filenames of input and output files via the DOS command line. It could also be run interactively, entering the commands at a prompt.

Distance 3.5 and later added a graphic user interface for defining the inputs. The program that does the actual work of analysis was called an "analysis engine", and was called D35Engine.exe in Distance 3.5, D4.exe in Distance 4, and now MCDS.exe. This program is run from the Distance graphical interface in batch mode. The exact way that Distance communicates with the MCDS analysis

engine is outlined in another Appendix – see Introduction to Inside Distance Appendix.

The data format and command language used to run MCDS.exe are therefore very similar to those used to run the old versions of Distance (the major differences are outlined in a subsection, below). The last complete documentation for the command language is the Distance 2.2 users manual, which is available for download from the support page of the Program Distance web site. Many new features have been added since Distance 2.2 (for example multiple covariates and flat data file input), but some features are also no longer supported. These include: interactive mode (batch mode only is now supported) and hierarchical data input (flat files only). For a full list, see the section Changes in MCDS Engine Since Distance 2.2.

Running the MCDS engine

The command to run the MCDS engine is

```
MCDS Parameter1, Parameter2
```

where

`Parameter1` is either a 0 or a 1. 0 is for run mode - i.e. run the analysis. 1 is for import mode, which is used to implement part of the Project Import feature in Distance and is not described further here.

`Parameter2` is the filename of the input command file – see MCDS Command Language, below for details of the contents of this file.

The program returns a number to the command line, indicating the status of the run, as well as up to 6 files of output – see Output From the MCDS Engine.

Example 1:

Assume that we have a command window open in the Distance program directory (usually `C:\Program Files\Distance5`), and that we have a file `TestInput.txt` in that directory. Then we type:

```
MCDS 0, TestInput.txt
```

Example 2:

Assume we have a command window open in some arbitrary directory (e.g., `C:\`). Assume that we have an input command file `C:\Temp\Input File.txt` that we want to run. Assume that the `MCDS.exe` program is in the Distance program directory `C:\Program Files\Distance5`. Because both the input file and the Distance program directory have spaces in them, we need to enclose the program and file names in quotes:

```
"C:\Program Files\Distance5\MCDS" 0, "C:\Temp\Input File.txt"
```

The space between MCDS and parameter 1, and the space and comma between parameter 1 and parameter 2 are critical. For example, in Example 1, above,
`MCDS 0,TestInput.txt`
will not work (it will return the value 4 - file error) because there is no space between parameter 1 and parameter 2 (see MCDS engine command line output for more about the numbers returned to the command line).

You can copy the file MCDS.exe to another folder and run it from there if you want to (e.g., C:\temp). You could also add the Distance program folder to your windows path (in Windows XP it's under Control Panel | System |

Advanced | Environment variables) so you don't then need to give the full path when calling it from the command line.

An example of how to run the MCDS analysis engine from a program written in another language (in this case Visual Basic) is given on the Support page of the Program Distance Web Site. See also the archives of the Distance-sampling Email List for some messages discussing how to do this (try searching for "stand alone").

MCDS Command Language

The MCDS analysis engine is driven by a relatively simple command language. When you run the MCDS engine from the command line (see Running the MCDS Engine), you pass in the name of a file containing these commands. The commands are then interpreted by the engine, which performs the analysis you have asked for.

The command file is divided into 4 sections:

- The Header section, where the location of the output files is specified. This is always 6 lines long.
- The Options section, where general program options are set. This begins with the `OPTIONS` command and ends with an `END` command.
- The Data section, where the location and format of the data file is specified. This begins with the `DATA` command and ends with an `END` command.
- The Estimate section, where estimation options are set. This begins with the `ESTIMATE` command and ends with an `END` command.

The format of each section is described in the following pages, and an example command file is given below. You can see many other examples by looking in the log tab of CDS or MCDS analyses that have been run.

The language interpreter is case insensitive. All commands (apart from the files in the header) end with a semicolon. Each command is usually given a new line, but this is not necessary. Commands and options can in theory be shortened so that they are the minimum length necessary to make them uniquely distinguishable – but this is not recommended as it leads to incomprehensible command files! The order of the commands in the Options, Data and Estimate sections should not matter.

An easy way to generate a template command file for a particular analysis is to set up that analysis using the Distance graphical interface, and then run the analysis in **Debug mode**. In this mode, the Distance interface generates a command file and data file, and stores them in the Windows temporary folder, but does not run the analysis. For more about Debug mode, see the Program Reference page on the Analysis Preferences Tab.

```
C:\Temp\dst111.tmp
C:\Temp\dst110.tmp
C:\Temp\dst112.tmp
C:\Temp\dst113.tmp
None
None
Options;
Type=Line;
```

```
Length /Measure='Mile';
Distance=Perp /Measure='Foot';
Area /Units='Square mile';
Object=Single;
SF=1.0;
Selection=Sequential;
Lookahead=1;
Maxterms=5;
Confidence=95;
Print=Selection;
End;
Data /Structure=Flat;
Fields=STR_LABEL, STR_AREA, SMP_LABEL, SMP_EFFORT, the MRDS engine;
Infile=C:\Temp\dst10D.tmp /NoEcho;
End;
Estimate;
Distance /Intervals=0,1,2,3,4,5,6,7,8 /Width=8 /Left=0;
Density=All;
Encounter=All;
Detection=All;
Size=All;
Estimator /Key=HN /Adjust=CO /Criterion=AIC;
Monotone=Strict;
Pick=AIC;
GOF;
Cluster /Bias=GXLOG;
VarN=Empirical;
End;
```

Example command file

Header Section

This section is required in all command files. Here, you specify the names of the output files that Distance will generate. If the files do not exist, they will be created. If they exist, they will be overwritten. The section is 6 lines long, and each line corresponds with the following file:

- Output file
- Log file
- Stats file
- Plot file
- Bootstrap file
-

Bootstrap progress file

If you are not using the bootstrap to estimate variance, you can specify "None" in the bootstrap and bootstrap progress file lines.

For information about the contents of these files once an analysis has run (or while the analysis is running in the case of the bootstrap progress file), see the section Output From the MCDS Engine.

Example:

```
C:\temp\dst6FA1.tmp
C:\temp\dst6FA0.tmp
C:\temp\dst6FA2.tmp
C:\temp\dst6FA3.tmp
None
None
```

If you do not include a path for the files (e.g., just `dst6FA1.tmp` in the above, for the first file), it is created and written into the current working directory (the directory you called the program from).

Note! In previous versions of Distance, the CDS and MCDS engine required 5 header lines, and not six (because there was no bootstrap progress file). Also, the bootstrap file came before the plot file. So, if you have any code for calling previous versions, you'll need to update it to call the new version.

Options Section

Various options can be set to control program operation. Once an option value has been set, it retains its value until you change it or exit the program. The data options define the characteristics of the data collected and how they are to be entered. The model fitting options define values to be used in fitting a probability density function to the distance data, some of which can be overridden in the estimation procedure. Print options control the amount and format of program output and bootstrap options control the number of bootstrap samples and the random number seed used to generate a bootstrap sequence.

This section should always begin with the command OPTIONS and end with the END command.

Below are the valid commands in the options section by category. Each option and its possible values are individually described in the following sections in alphabetical order.

Miscellaneous	
DEFAULT command	Options reset to default
END command	Ends options section
LIST command	List option values
Output	
PRINT command	Controls amount of output
QQPOINTS command	Max number of points in qq plot
TITLE command	Value of output title
Data Options	
AREA command	Set area quantities
CUERATE command	Set cue rate
DISTANCE command	Set distance quantities
LENGTH command	Set length quantities
OBJECT command	SINGLE or CLUSTER
SF command	Sampling fraction
TYPE command	POINT, LINE or CUE
Model Fitting	
LOOKAHEAD command	Max for sequential fit
MAXTERMS command	Max # model parameters
PVALUE command	Significance level (α-level)
SELECTION command	Term selection mode
Model Fitting	
BOOTSTRAPS command	# of bootstrap samples
SEED command	Random number seed

AREA Command

Syntax:

AREA /CONVERT=value /UNITS='label' ;

Description:

This command defines the area unit for expressing density (D). The switches are:

/UNITS='label' - a label for the unit of area of the density estimate. The single quotes are only required to retain lowercase. Only the first 15 characters are used.

/CONVERT=value - value specifies a conversion factor which is used to convert the estimated density to new units for area. It is needed for atypical units.

If the MRDS engine recognizes the measurement unit for DISTANCE (and LENGTH for line transects) and if it recognizes the Area UNITS label, it will calculate the appropriate conversion factor. However, if one or more of the UNITS is not recognized, you will need to specify the conversion value with the CONVERT switch. The Area units recognized by the program are those listed under the DISTANCE command and HECTARES (HEC) and ACRES (ACR). For example, the unit can be entered as Squared Meters or Metres Squared because the MRDS engine recognizes the unit based on the character string MET. See the the MRDS engine command below for a definition of recognized units

Default: AREA /UNITS=HECTARES;

Examples:

Distances are measured in feet but analyzed in meters, length is measured in miles and density is estimated as numbers per square kilometer. The MRDS engine will do necessary unit conversions because all unit labels are recognized.

DISTANCE /MEASURE=FEET /UNITS='Meters';

LENGTH/UNITS='Miles';

AREA /UNITS='Sq. kilometers';

BOOTSTRAPS Command

Syntax:

BOOTSTRAPS=value ;

Description:

"Value" is the number of bootstrap samples which should be generated. For a reasonable variance estimate, this number should be at least 100. We recommend setting BOOTSTRAPS=999 or 1000 to construct a bootstrap confidence interval.

Default: BOOTSTRAPS=1000;

CUERATE Command

Syntax:

CUERATE = value1 /SE=value2 /DF=value3;

Description:

For cue counting, "value1" is the average rate at which animals issue visual or auditory detection cues. The rate should be given in the same units of time as the values given for sampling effort in the data. For example, if effort is measured in hours then the cue rate should be number of cues per hour. The cue

rate must be a positive number (>0). Optionally a standard error for the cue rate can be given with "value2", and the degrees of freedom can be given with "value3" (a DF of 0.0 is interpreted as infinite degrees of freedom). The standard error and df is accounted for in the estimated standard error of the density and abundance estimates. This option is only used if TYPE=CUE is specified.

Default: CUERATE=1 /SE=0 /DF=0;

Example:

An estimate of the cue rate is 12 per hour with a standard error of 2 per hour and 93 degrees of freedom. The sample effort for this cue counting example be specified in hours sampled.

CUERATE=12 /SE=2 /DF=93;

DEFAULT Command

Syntax:

DEFAULT ;

Description:

This command resets all of the options to their default values. Remember that an option remains in effect until it is changed or the MRDS engine is terminated. The default values for each of the options are:

$PVALUE = 0.15$	$PRINT = SELECT$	$TYPE = LINE$
$SELECT = SEQUENTIAL$	$SQUEEZE = OFF$	$OBJECT = SINGLE$
$MAXTERMS = 5$	$BOOTSTRAPS = 1000$	$DISTANCE = PERP / EXACT / UNITS = METERS$
$ITERATIONS = 100$	$SEED = 0$	$LENGTH / UNITS = KILOMETERS$
$LOOKAHEAD = 1$	$SF = 1$	$AREA / UNITS = HECTARES$
$CUERATE = 1 / SE = 0 / DF = 0$		

DISTANCE Command

Syntax:

$$DISTANCE = \left| \begin{matrix} PERP \\ RADIAL \end{matrix} \right| \left| \begin{matrix} / WIDTH = width \\ / CONVERT = value \\ / MEASURE = 'label' \\ / LEFT = left \end{matrix} \right| \left| \begin{matrix} / NCLASS = nclass \\ / UNITS = 'label' \\ / RTRUNCATE = t \\ / INTERVALS = c_0, c_1, \ldots, c_u \\ / EXACT \end{matrix} \right| ;$$

Synonyms: RIGHT=WIDTH

Description:

This command describes numerous features about the distance data and defines the default values for estimation. The format of the data entry within the Data section is determined by the values set with this command. Whereas, the DISTANCE command in the Estimate section only determines how the distance data are analyzed.

For line transect data (TYPE=LINE), this command defines whether the data will be entered as either perpendicular distances or as radial distance and angle measurements.

- **PERP** - perpendicular distance was measured for a line transect
- **RADIAL** - radial distance and angle were measured in line transects

For TYPE=POINT (which includes trapping webs) or CUE, the MRDS engine=RADIAL is assumed and only radial distances are expected.

Distances can be entered as ungrouped or grouped. Ungrouped implies an exact distance is entered for each observation in the data. Grouped means a set of distance intervals is given and the frequency of observations in each interval is entered. Ungrouped distances are indicated by the switch /EXACT and grouped data is indicated by the /INTERVALS switch which also specifies the distance intervals (c0-c1,c1-c2,c2-c3,...). The value c0 specifies the left-most distance and cu the right-most distance for grouped data. Typically, c0=0 and cu=w. Intervals can also be specified by using the /NCLASS and /WIDTH and optionally the /LEFT switch. These switches will create 'nclass' equal width distance intervals between the values of 'left' and 'width' (i.e., each interval is of length (width-left)/nclass). For ungrouped data, it is also possible to specify left and right truncation with the /LEFT and /WIDTH switches. Any values outside of these bounds are excluded from the analysis. Right truncation as a percentage of the observations can also be specified for both grouped and ungrouped data with /RTRUNCATE switch. The value of t must be between 0 and 1. In the analysis, no more than t*100% of the data is truncated from the right. For ungrouped data, the width is set at the distance which represents the (1-t)*100% quantile. For grouped data, intervals are truncated from the right as long as no more than t*100% of the data is truncated. If t=0 and the data are ungrouped data, the width is set to the largest distance measurement and if the data are grouped, the width is set to the endpoint for the right-most interval with a non-zero frequency. For ungrouped data, if both the /WIDTH and /RTRUNCATE switch are specified, the RTRUNCATE value specifies the value of width.

The DISTANCE command is also used to define the measurement unit for distances:

/MEASURE = 'label' - a label for the units in which distance was measured. Single quotes are only required to retain lowercase. Only the first 15 characters are used.

/UNITS='label' - a label for the units for distance after conversion, if any. Single quotes are only required to retain lowercase. Only the first 15 characters are used.

/CONVERT=value - value specifies a conversion factor which is used to convert the input distances for atypical units.

MEASURE and UNITS switches are used to convert from the unit in which the data are recorded and entered (MEASURE) to the unit for analysis (UNITS). It is not necessary to convert distances to different units for analysis as long as it is a unit that is recognized by the MRDS engine (see list below). It is only provided as a convenience and it is probably easier to leave measurements in their original units. If you do convert units, take note that values such as f(0), h(0), effective strip width (ESW) and effective detection radius (EDR) are expressed in the converted units. Thus, the point estimate and standard errors will change by the conversion factor from the measured to analysis units. If you are not converting distance units, you can specify the units with either switch (/MEASURE or /UNITS). The most common measurement units are recognized by the MRDS engine and there is no need to enter a conversion value (/CONVERT= value).

The following are the recognized measurement unit labels:

- CENTIMETERS
- METERS
- KILOMETERS
- MILES

- INCHES
- FEET
- YARDS
- NAUTICAL MILES

Each label is recognized by its first 3 characters which allows variations in spelling. For example, if you enter METRES, it will use METRES as the label and will recognize it based on MET. Values are given in uppercase but can be entered in upper or lowercase. If the MRDS engine recognizes the /UNITS and /MEASURE labels and you specify the /CONVERT= switch it will display a warning message that you are overriding the conversion value. Values for /WIDTH, /LEFT, and /INTERVALS should be given in original measurement units and not in converted units.

Default:

DIST=PERP /UNITS='Meters' /MEASURE='Meters' /EXACT /LEFT=0 /RTRUNCATE=0;

Examples:

Perpendicular distance measured in intervals of 2 feet to a distance of 10 feet and converted to metres (meters) for analysis. The grouped data are entered as the frequency of observations in each of the 5 distance intervals (see the Data section). Notice that WIDTH is specified in the original measurement units of feet and not in meters.

DIST=PERP /MEASURE='Feet' /UNITS='Metres' /WIDTH=10 /NCLASS=5 ;

LENGTH Command

Syntax:

LENGTH /CONVERT=value /UNITS='label' /MEASURE='label' ;

Description:

This command sets the measurement unit for line length and any desired conversion to different units for analysis. It is not necessary to convert line length, but may be desirable depending on the original units.

/MEASURE='label' - a label for the units in which line length was measured. Single quotes are only required to retain lowercase. Only the first 15 characters are used.

/UNITS='label' - a label for the units for length after conversion, if any. Single quotes are only required to retain lowercase. Only the first 15 characters are used.

/CONVERT=value - value specifies a conversion factor which is used to convert length measured in atypical units.

See further explanation under the DISTANCE command for the /MEASURE, /UNITS and /CONVERT switches. The LENGTH command is used for line transects only.

Default:

LENGTH /UNITS=KILOMETERS /MEASURE=KILOMETERS;

Example:

Length is entered in miles but converted to kilometers for display and analysis.

LENGTH /UNITS='Kilometers' /MEASURE='Miles' ;

LIST Command

Syntax:

LIST;

Description:

Lists current values of the program options and the program limits to the screen.

LOOKAHEAD Command

Syntax:

LOOKAHEAD=value ;

Description:

For term selection modes SEQUENTIAL and FORWARD (see SELECTION command), "value" specifies the number of adjustment terms which should be added to improve the fit, before the added terms are considered to be non-significant. For example, if LOOKAHEAD=2 and a model with 2 adjustment terms does not significantly improve the fit over a model with 1 term, a model with 3 adjustment terms is fitted. If the 3-term model is an improvement over a 1-term model, the algorithm will continue with the 3-term model as the new base model. If it is not an improvement, the 1-term model would be chosen. If LOOKAHEAD=1 (the default), in the above example, the 3-term model would not have been examined because upon finding the 2-term model was not an improvement, the 1-term model would have been used.

Default: LOOKAHEAD=1;

MAXTERMS Command

Syntax:

MAXTERMS=value ;

Description:

"Value" is the maximum number of model parameters. The maximum number of adjustment terms (defined as *m*) that may be added is MAXTERMS minus the number of parameters in the chosen key function (defined as *k*). MAXTERMS must be less than or equal to 5. This option is only useful to limit the number of model combinations with the term selection mode that considers also possible combinations of adjustment terms (SELECTION=ALL). Use the NAP switch on the Estimator command to specify an exact number of adjustment terms to be used. The maximum number of adjustment terms is also limited by the number of observations for ungrouped data or number of distance intervals for grouped data.

Default: MAXTERMS=5;

OBJECT Command

Syntax:

$$OBJECT = \left| \begin{matrix} SINGLE \\ CLUSTER \end{matrix} \right| ;$$

Description:

This option defines whether objects are detected individually (SINGLE) or as clusters (CLUSTER).

SINGLE - Object always detected as a single animal or other entity (e.g., duck nest)

CLUSTER - Object detected as a cluster (e.g., herd, flock,pod of whales)

Default: OBJECT=SINGLE;

PRINT Command

Syntax:

$$PRINT = \left| \begin{array}{c} SELECTION \\ RESULTS \\ ALL \\ SUMMARY \end{array} \right| ;$$

Description:

This option sets the default level of printing in the output. The various settings are hierarchical and more control over the amount of results can be obtained with the PRINT command in the Estimate section.

ALL - print fitting iterations, model selection results and estimation results

SELECTION - print model selection results and estimation results

RESULTS - print estimation results only

SUMMARY (NONE) - only summary tables are printed

Note: if you choose RESULTS or SUMMARY, warnings are not given about the algorithm having difficulties fitting a particular model or constraining the fit to achieve monotonicity.

Default: PRINT=SELECTION;

PVALUE Command

Syntax:

PVALUE=α;

Description:

α is the significance level of likelihood ratio tests to determine significance of adding adjustment terms and is the default value for the significance test for size-bias regression of cluster sizes

Default: PVALUE=0.15;

QQPOINTS Command

Syntax:

PVALUE=value;

Description:

Maximum number of points to print in qq-plots. When there are a large number of data points, plotting all the points can take quite a while and result in a very large plot file. The default, 0, means no maximum – i.e., plot every point.

Default: QQPOINTS=0;

SEED Command

Syntax:

SEED=value ;

Description:

SEED specifies the random number seed for generating a sequence of random numbers for bootstrap samples. "Value" should be a large odd number preferably greater than 2,000,000. If you use the same seed, the same sequence

of random numbers will be generated. You can use SEED=0; (the default) which will use a value from the computer's clock to generate a seed.

Default: SEED=0;

SELECTION Command

Syntax:

$$SELECTION = \begin{matrix} SEQUENTIAL \\ FORWARD \\ ALL \\ SPECIFY \end{matrix} ;$$

Description:

This command specifies the default mode for adjustment term selection in the ESTIMATE procedure for fitting the detection function. The /SELECT switch of the ESTIMATOR command overrides the default value. See the Estimate section for a description of adjustment term and model selection.

SEQUENTIAL - add adjustments sequentially (e.g., for simple polynomial, in increasing order of the exponent).

FORWARD - equivalent to forward selection in regression; select the adjustment term which produces the largest increase in the maximum of the likelihood function

ALL - fit all combinations of adjustment terms with the key function and use the model with smallest Akaike Information Criterion (AIC) value.

SPECIFY - user-specified number of adjustment terms and possibly order of the adjustments.

Default: SELECTION=SEQUENTIAL;

SF Command

Syntax:

SF=c ;

Description:

SF defines the value of the sampling fraction which is typically 1. However, if only one side of a transect line is observed c=0.5, or if some fraction of the circle surrounding a point transect is searched, c is the fraction searched (e.g.,c=0.5 if a semi-circle is observed). For cue counting, c is the proportion of a full circle that is covered by the observation sector. For a sector of 90^o (45^o either side of the line) with cue counting, c = 0.25

Note that SF can now be specified using the MULTIPLIER command, with SE=0, and this is the way that Distance does it.

Default: SF=1;

TITLE Command

Syntax:

TITLE='yourtitle' ;

Description:

This command sets a value for the title which is printed at the top of each page. Yourtitle should contain no more than 50 characters. Excess characters are not used. There is only 1 title line. Re-specifying the title will replace the previous value.

TYPE Command

Syntax:

$$TYPE = \begin{vmatrix} POINT \\ LINE \\ CUE \end{vmatrix} ;$$

Description:

This option defines the type of sampling, which determines what types of data can be entered and how data are analyzed.

POINT - point transect data

LINE - line transect data

CUE - cue counting data

Trapping webs should be treated as point transects.

Default: TYPE=LINE;

Data section

In this section, you specify the file containing data, and which column in this file corresponds with which field. For more about the format of the data file, see see the section describing the MCDS Engine Required Data Format.

The data section should always begin with the statement DATA /STRUCTURE=FLAT and end with the statement END. Note that historical versions of this engine used a hierarchical data format, but that is no longer supported, so the /STRUCTURE=FLAT switch is now mandatory.

The commands that are valid in the Data section are listed in alphabetical order below, and described in the following sections.

Data section commands	
END command	Ends data section
FACTOR command	Specifies that a field is a factor covariate
FIELDS command	List of fields in the data file
INFILE command	Gives filename of data file
SIZEC command	Specifies that a field is the cluster size covariate

FACTOR command

Syntax:

FACTOR /NAME='fieldname' /LEVELS=value /LABELS= 'label1', 'label2', … ;

Description:

This command defines a field in the data file as a factor covariate in MCDS analyses. For more about factor covariates, see the section on factor and non-factor covariates in MCDS in Chapter 9 of the Users Guide. Covariates for the detection function are specified in the Estimator command.

There should be one FACTOR command for each factor field in the data file. If there are no factor fields, this command will not be present.

/NAME='fieldname' – the name of the field (must be one of the names in the FIELDS command)

/LEVELS=value – the number of levels in the factor covariate

/LABELS= 'label1', 'label2', … - a comma-delimited list giving the value of each level of the factor

Default: no FACTOR command

Examples:

The data file contains a column for obsererver, which is specified in the FIELDS command as "Observer". Observer is to be used as a factor covariate in the detection function, and each observation can take one of three possible values "Peter", "Paul" and "Mary".

FACTOR /NAME=Observer /LEVELS=3 /LABELS=Peter, Paul, Mary;

FIELDS command

Syntax:

FIELDS= fieldname1, fieldname2, fieldname3, …

Description

This command gives a list of the fields occuring in the data file, reading the columns of the data file from left to right. The following fieldnames are required:

- SMP_LABEL – sample label
- SMP_EFFORT – sample effort (line length/number of points)
- DISTANCE – perpendicular or radial distances (depending on the TYPE and DISTANCE commands)

If the TYPE = LINE and DISTANCE = RADIAL then an additional required field is

- ANGLE – angle of radial distances, in degrees

If OBJECT = CLUSTER then another required field is

- SIZE – the cluster size

Two additional fields with fixed names may be specified:

- STR_LABEL – stratum label
- STR_AREA – stratum area – if areas are ommitted then density but not abundance is calculated

If covariates are specified in the ESTIMATOR command then these should be included in the data file, and their names listed in the FIELDS command. In addition to being listed in the FIELDS command, factor covariates should be declared as such using a FACTOR command.

Default: No default

Examples:

Standard line transect data, with a column for stratum label, area, transect label, line length and perpendicular distance:

Fields=STR_LABEL, STR_AREA, SMP_LABEL, SMP_EFFORT, DISTANCE

Line transect data, with radial distance and angle, objects as clusters, and an additional field for an Observer covariate:

Fields=STR_LABEL, STR_AREA, SMP_LABEL, SMP_EFFORT, DISTANCE, ANGLE, SIZE, Observer

INFILE Command

Syntax:

INFILE = filename	*/ ECHO* */ NOECHO*	;

Description:

This command specifies the data file name. Filename should either give the full absolute path to the file, or just the filename if the file is in the current directory.

The ECHO and NOECHO switches control whether the data are ECHOed to the LOG file. Once you are certain that the data are free of errors, using /NOECHO will reduce the amount of output to the LOG file.

Example:

Infile=C:\temp\dst7035.tmp /NoEcho;

SIZEC command

Syntax:

SIZEC=fieldname;

Description

Specifies that the field "fieldname" is the cluster size field, when cluster size is a covariate in the detection function.

Estimate section

The following are valid commands in the estimate section:

Estimate section commands	
BOOTSTRAP command	bootstrap variance/confidence intervals
CLUSTER command	estimation of expected cluster size
DENSITY command	resolution of density estimation
DETECTION command	resolution of detection probability estimation
DISTANCE command (Estimate section)	analysis treatment of distances
ENCOUNTER command	resolution of encounter rate estimation
END command	initiates estimation
ESTIMATOR command	model for g(x)
G0 command	estimate of g(0) and its standard error
GOF command	intervals for goodness of fit test/display
MONOTONE command	monotonicity constraints on g(x)
MULTIPLIER command	multipliers in the detection function
PICK command	method of model choice
PRINT command (Estimate section)	detailed control of output
SIZE command	resolution of expected cluster size estimation
VARN command	variance estimation of n

The commands are described below in alphabetical order. You will use these commands to define:

- which quantities you want to estimate and at what level of resolution (DENSITY, DETECTION, ENCOUNTER, SIZE),
- how distance and cluster size are treated in the analysis and which models are used for estimation (DISTANCE, CLUSTERS, ESTIMATOR, MONOTONE, PICK, GOF, G0),

- how variances are estimated (VARN, VARF, BOOTSTRAP), and
- how much output should be generated (PRINT).

Density and abundance estimates are comprised of the following components:

- detection probability,
- encounter rate,
- expected cluster size (if the detected objects are clusters).

It is possible to restrict estimation to one or more of these components without estimating density; however, all components must be estimated to obtain an estimate of density. You will use the commands DENSITY, DETECTION, ENCOUNTER, and SIZE to define which components will be estimated. If you do not use any of these commands, each component and density is estimated, by default. Likewise, if you use the DENSITY command, density and all of its components are estimated. If you use any or all of the DETECTION, ENCOUNTER, and SIZE commands and not the DENSITY command, only the specified components are estimated. For example,

```
ESTIMATE;
  ENCOUNTER ALL;
END;
```

will only estimate encounter rate.

Estimates of density and its components can be made at different levels of the sampling hierarchy (Sample < Stratum < All). The DENSITY, DETECTION, ENCOUNTER, and SIZE commands are used to specify the level at which each quantity is estimated. Different levels can be used for the various quantities; although, some combinations are incompatible. An error message is given, if the levels are incompatible. The lowest level of resolution specified for DENSITY is the default level for each of its components, if they are unspecified. For example,

```
ESTIMATE;
  ESTIMATOR /KEY=UNIFORM;
  DENSITY BY STRATUM;
END;
```

will estimate density and each of its components for each stratum defined in the data. The lowest level for density must coincide with a level assigned to encounter rate. The level of any component cannot be lower then the lowest level specified for density. For example, the following is not valid:

```
ESTIMATE;
  ESTIMATOR /KEY=UNIFORM;
  DENSITY BY STRATUM;
  DETECTION BY SAMPLE;
END;
```

If a size-bias regression estimate of expected cluster size is computed, the level for SIZE must be no greater than the level for DETECTION. This feature is most useful for estimating density by stratum when too few observations exist in each stratum to estimate f(0) (or h(0)). A solution is to assume f(0) is the same for all strata, which is illustrated in the following example:

```
ESTIMATE;
  ESTIMATOR /KEY=UNIFORM;
  DENSITY BY STRATUM;
  DETECTION ALL;
END;
```

All of the observations are pooled to estimate a common value for f(0), which is used in each stratum density estimate.

When there are covariates in the detection function, it is possible to fit the detection function at one level, and then estimate probability of detection at a lower level. For more on this, see Chapter 9 of the Users Guide Estimating the detection function at multiple levels. An example of this, with a global detection function fit with a habitat covariate which is specific to each stratum, and then the detection function estimated by stratum is:

```
ESTIMATE;
  ESTIMATOR /KEY=UNIFORM /COVARIATES=Habitat;
  DENSITY BY STRATUM;
  DETECTION ALL;
  DETECTION BY STRATUM
END;
```

Possibly the most confusing aspect of estimation with the MRDS engine will be the specification of models for detection probability and model selection. A model is specified with the ESTIMATOR command which defines a type of key function and adjustment function. The adjustment function is actually a series of terms which are added to the key function to adjust the fitted function to the data. Model selection includes 1) selecting how many and which adjustment terms are included in the model (term selection) and 2) selecting a "best" model (estimator) from the specified set of competing models.

The default method of selecting terms (term selection mode) is defined by the SELECTION command in the Options section. Its value can be overridden with the /SELECT switch of the ESTIMATOR command. Related options include LOOKAHEAD and MAXTERMS. There are 4 types of term selection modes described below: 1) SEQUENTIAL, 2) FORWARD, 3) ALL, and 4) SPECIFY. The maximum number of adjustment terms that can be included in the model is limited by the value of (MAXTERMS - number of parameters in the key function) and less frequently by either the number of observations, for ungrouped data, or the number of distance intervals for grouped distance data. The MRDS engine will issue a warning message if the number of parameters is being limited by the amount of data.

Term selection mode SPECIFY implies the user will specify which adjustment terms are included in the model. Typically, this is used to specify that a key function without adjustment terms is to be fitted to the data, as in the following example:

```
ESTIMATE;
  ESTIMATOR /KEY=HNOR /NAP=0 /SELECT=SPECIFY;
END;
```

It is not necessary to include the /SELECT switch, but it will prevent the MRDS engine from issuing a warning message that you are specifying the model. It is also possible to specify any combination of terms and give starting values for their coefficients. the MRDS engine does not select the terms to include in the model but does estimate the parameters to fit the model to the data. For example,

```
ESTIMATE;
  ESTIMATOR /KEY=UNIFORM /NAP=2
    /SELECT=SPECIFY /ORDER=1,3;
END;
```

specifies the model as the following 2-term cosine series for which the parameters a1 and a2 are estimated:

$$f(x) = \frac{1}{w}\left(1 + a_1 \cos\left(\frac{\pi x}{w}\right) + a_2 \cos\left(\frac{3\pi x}{w}\right)\right)$$

ALL, as its name implies, examines all possible combinations of a limited number of adjustment terms. If z is the maximum number of parameters

(MAXTERMS=z) and k is the number of parameters in the key function, then there are 2z-k combinations of the adjustment terms. Each model is fitted to the data and the model with the smallest value of the Akaike's Information Criterion (AIC) is selected.

SEQUENTIAL and FORWARD both consider a subset of models with different combinations of adjustment terms. For each of these term selection modes, a sequence of models is considered. An adjustment term is added at each step of the sequence. The sequence of models can be represented as:

M1- key function with no adjustment terms
M2- key function with 1 adjustment term
M3- key function with 2 adjustment terms
:
:
Mv- key function with v-1 adjustment terms

A stopping rule (CRITERION) is either based on a likelihood ratio test or minimizing AIC. Model Mt is chosen if there is no model in the sequence Mt+1,...,Mt+l which provides a significantly better fit as determined by the specified CRITERION. The LOOKAHEAD option determines the length (l) of the sequence of models that is examined before choosing model Mt. SEQUENTIAL and FORWARD only differ in their choice of which adjustment term is included at each step in the sequence. SEQUENTIAL term selection adds the terms sequentially based on the order of the term. For polynomial adjustment functions, the order of the adjustment term is the exponent of the polynomial. Terms are added in the following sequence: $(x^t, x^{t+2}, \ldots$. For cosine adjustments, cosine terms are added in the following sequence: $\cos(t\pi x/w), \cos((t+1)\pi x/w), \ldots$. The beginning value, t, is determined by the shape of the key function. FORWARD selection adds 1 term at a time but not necessarily in sequential order. For each model in the sequence, each term not already in the model is added and the adjustment term which increases the likelihood the most is chosen as the term to add. For example, to find model M2, z-k models are fitted to the data, each with a single adjustment term of a different order (e.g., $x^2, x^4, x^6, x^8, or\ x^{10}$). The term which maximizes the likelihood is selected for model M2. Model M3 would then consider adding another term not included in M2. With FORWARD selection it is possible to select models that cannot be selected with the SEQUENTIAL mode. For example, the following model might be chosen with FORWARD selection:

$$f(x) = \frac{1}{w}\left(1 + a_1 \cos\left(\frac{\pi x}{w}\right) + a_2 \cos\left(\frac{3\pi x}{w}\right)\right)$$

However, with SEQUENTIAL selection, the adjustment term $\cos\left(\frac{3\pi x}{w}\right)$, could not be added without first adding the adjustment term $\cos\left(\frac{2\pi x}{w}\right)$.

The additional level to model fitting is to choose between the competing models (ESTIMATORs). This model selection step is determined by the PICK command. It has 2 values: NONE and AIC. If you assign the value NONE, the MRDS engine will not choose between the different models and will report the estimates for each model. However, if you accept the default value, AIC the MRDS engine will only compute estimates based on the model which has the smallest AIC value.

BOOTSTRAP Command

Syntax:

BOOTSTRAP	*/ STRATUM* */ INSTRATUM*	*/ SAMPLES* */ OBS*	;

Description:

The BOOTSTRAP command initiates a non parametric bootstrap of the density estimation procedure. The number of bootstraps performed is determined by the BOOTSTRAP command in OPTIONS. The basic re-sampling unit of the bootstrap is a SAMPLE; however, if strata are replicates they can also be re sampled with the /STRATUM switch. If both are specified, re-sampling occurs at both levels (see example below). The switch /INSTRATUM can be set to restrict the re-sampling of samples or observations within stratum It would be used if density is estimated by stratum or sampling was stratified apriori.

The switch /OBS can be set to re-sample distances. Using BOOTSTRAP/OBS; will provide a non-parametric bootstrap of f(0) or h(0) and, if the population is clustered, E(s). However, the variances and confidence intervals are conditional on the sample size and do not include the variance of the encounter rate. The /OBS switch has been included for completeness but its routine use is not recommended. Reasonable confidence intervals for density could only be obtained by adding a variance component for the encounter rate. It is also possible to include the /OBS switch with /SAMPLES, however, this is not recommended unless the number of observations per sample is reasonable (> 15).

By default, issuing the BOOTSTRAP command without switches is equivalent to BOOTSTRAP/SAMPLES/INSTRATUM;. We recommend the default or dropping the /INSTRATUM if sampling across strata is appropriate. The use of /STRATUM is only appropriate if the strata represent an additional level of sampling (e.g., independent observers (stratum) traversing an independent set of line transects (sample)).

Each bootstrap resample is made up by sampling with replacement an equal number of units at the level you specify. For example, if you specify to resample samples within strata (the default), then each bootstrap resample is made up of the same number of samples (line or point transects) as your original sample, chosen randomly with replacement from the original sample (within each stratum). Note that for line transects, this means that the survey effort (total line length) will differ in each resample. Note also that each of your original samples has an equal probability of appearing in the resample (an alternative, which we do not implement, would be to have probability proportional to line length).

The bootstrap summary is given at the end of the output. The point estimate is the mean of the bootstrap point estimates. Two sets of confidence intervals are given: 1) log based confidence intervals based on a bootstrap standard error estimate, and 2) 2.5% and 97.5% quantiles of the bootstrap estimates (i.e., percentile confidence intervals).

Summary results from each iteration are stored in the Bootstrap file – see MCDS Engine Bootstrap File for details.

Example:

Re sample strata and samples within each stratum.

BOOTSTRAP /STRATUM /SAMPLES;

CLUSTER Command

Syntax:

$$CLUSTER \quad / WIDTH = value \quad / TEST = \alpha$$

$$\left| \begin{array}{c} / MEAN \\ / BIAS = \begin{array}{c} X \\ XLOG \\ GX \\ GXLOG \end{array} \end{array} \right|$$

Description:

The CLUSTER command, like the the DISTANCE command, is used to modify the way the cluster sizes are used in the estimate of density. By default, the WIDTH is chosen to match the truncation value set by the the MRDS engine command and the MRDS engine computes a size bias regression estimate (/BIAS=GXLOG) by regressing the loge(s) (natural logarithm specified as log() in the output) against (x), where x is the distance at which the cluster was observed.

The WIDTH switch specifies that only cluster sizes for observations within a distance less than WIDTH are used in the calculation of the expected cluster size. This treatment of the data can only be accomplished if the distances and cluster sizes are both entered as ungrouped.

The MEAN switch specifies that the expected cluster size is to be estimated as the average (mean) cluster size. Likewise, the BIAS switch specifies that expected cluster size is to be estimated by a size bias regression defined by the value of the switch.

Value	Meaning
X	Regress cluster size against distance x
XLOG	Regress loge(s) against distance x
GX	Regress cluster size (s) against (x)

The TEST switch specifies the value of the significance level to test whether the regression was significant. If it is non significant, the average cluster size is used in the estimate of density. The default value for the significance level is set by PVALUE in OPTIONS. If the TEST switch is not specified, the size bias regression estimate will be used regardless of the test value.

Examples:

Estimate the expected cluster size from the loge(s) vs. (x) regression, but use the average cluster size if the correlation is non significant as determined by the α-level set with PVALUE (default=0.15).

CLUSTER /BIAS=GXLOG /TEST;

DENSITY Command

Syntax:

$$\text{DENSITY by SAMPLE /DESIGN} = \left| \begin{array}{c} \text{NONE} \\ \text{REPLICATE} \end{array} \right| ;$$

or

DENSITY by STRATUM /DESIGN = NONE | STRATA | REPLICATE /WEIGHT = EFFORT | AREA | NONE ;

or

DENSITY by ALL;

Description:

These commands define the levels at which density estimates are made and how these estimates are weighted. If the DENSITY by ALL; command is used or if none of the commands (DENSITY, ENCOUNTER, DETECTION, SIZE) are given, all of the data are used to make one overall estimate of density.

If the DENSITY BY SAMPLE command is given, density is estimated for each sample. The DESIGN value defines how the estimates should be treated to create a pooled estimate. If DESIGN=REPLICATE (default), each sample is treated as an independent replicate from the stratum or the entire area. In this case, the estimates are weighted by effort (e.g., line length) to get a stratum density estimate (if DENSITY by STRATUM is also specified) or a pooled overall density estimate (see eqns. 3.84-3.87 in Buckland et al. (2001)). If DESIGN=NONE, the sample estimates are not pooled.

If DENSITY BY STRATUM is specified, an estimate is made for each stratum. A stratum estimate is a pooled estimate of the sample estimates within the stratum, if DENSITY by SAMPLE; is specified, or it is an estimate based on the data within the stratum. An overall (pooled) estimate of density is made unless DESIGN=NONE is specified. If DESIGN=REPLICATE, the stratum estimates are treated as replicates to create a pooled estimate and variance weighted by effort (eqns. 3.84-3.87 in Buckland et al. (1993), treating stratum as a sample).

If DESIGN=STRATA, the pooled estimate is a weighted sum of the estimates and the variance is a weighted sum of the stratum variances (Section 3.7.1 in Buckland et al. (2001)). Weighting is defined by the WEIGHT switch. If WEIGHT=NONE, the densities are summed, which is only useful if the population is stratified as by sex or age. If WEIGHT=AREA, the densities are weighted by area (which is the same as adding abundance estimates) and if WEIGHT=EFFORT, the densities are weighted by effort.

Prior to version 2.1, a combined estimate of abundance (N) was created by multiplying the combined density estimate by the sum of the areas specified on each of the STRATUM commands. This produces obviously erroneous results when DENSITY by STRATUM/DESIGN=REPLICATE; is used and the area size is repeated on each STRATUM. To avoid this problem in the following two situations the combined abundance estimate uses the area from the first stratum:

1) DENSITY by STRATUM/DESIGN=REPLICATE;

2) DENSITY by STRATUM/DESIGN=STRATA/WEIGHT=NONE;

In all other cases, the area is totalled from all of the strata.

Default:

For **DENSITY by SAMPLE**: **/DESIGN=REPLICATE**

For **DENSITY by STRATUM**: **/DESIGN=STRATA /WEIGHT=AREA**

Examples:

An estimate is needed for each stratum and it will be weighted by stratum area:

DENSITY BY STRATUM;

An estimate is needed for each stratum and there are enough observations in each sample to get an estimate from each. The strata represent different platforms surveying the same area so the strata are treated as replicates.

DENSITY BY SAMPLE;
DENSITY BY STRATUM /DESIGN=REPLICATE;

DETECTION Command

Syntax:

DETECTION by | SAMPLE / STRATUM / ALL |

Description:

This command explicitly specifies that detection probability (and its functionals f(0), h(0)) should be estimated and the resolution at which the estimate(s) should be made (by SAMPLE, by STRATUM, or ALL data).

When there are covariates in the detection function, it is possible to fit the detection function at one level, and then estimate probability of detection at a lower level. For more on this, see Chapter 9 of the Users Guide Estimating the detection function at multiple levels.

Examples:

Density is estimated by stratum but the estimates are based on an estimate of f(0) for all the data.

DENSITY by STRATUM;
DETECTION ALL;

Estimate detection by stratum choosing between 2 models but do not estimate any other parameters. Different models may be selected for each stratum.

ESTIMATE;
DETECTION BY STRATUM;
ESTIMATOR/KEY=HAZ;
ESTIMATOR/KEY=UNIF;
END;

Fit detection function globally using a habitat covariate, but estimate by stratum (the level at which habitat is defined):

ESTIMATE;
ESTIMATOR /KEY=UNIFORM /COVARIATES=Habitat;
DENSITY BY STRATUM;
DETECTION ALL;
DETECTION BY STRATUM
END;

DISTANCE command (Estimate section)

Syntax:

DISTANCE /WIDTH=value /INTERVALS=c_0, c_1, … , c_u /LEFT=l /RTRUNCATE=t /NCLASS=nclass /SMEAR=angle,pdist;

Description:

The DISTANCE command is used to specify the way the distances should be treated in the analysis. It can be used to specify left and right truncation of the distance, to group the distances into intervals prior to analysis, and to "smear" radial distance / angle measurements into perpendicular distance intervals.

Right truncation is specified with either the WIDTH or RTRUNCATE switch. If WIDTH=w, only distances less than or equal to w, are used in the analysis. If RTRUNCATE=t, the right truncation distance is set to use (1-t)*100% of the data. If the data are ungrouped the (1-t)*100 percentile is used as the truncation distance. If the distances are grouped or being analyzed as such, the truncation distance is set to the uth interval end point where u is the smallest value such that no more than t*100% of the distances are truncated. The value of t=0 trims the intervals to the right most interval with a non-zero frequency. If both the WIDTH and RTRUNCATE are specified, the value of RTRUNCATE defines the truncation unless the WIDTH is used with NCLASS (see below).

Left truncation is accomplished with the LEFT switch which works in an analogous fashion to WIDTH. If LEFT=l, only distances greater than or equal to l are used in the analysis. If LEFT is not specified, it is assumed to be 0.

The INTERVALS command is used to specify u distance intervals for analyzing data in a grouped manner when the data were entered ungrouped. The value c0 is the left most value and so it can be used for left truncation. If there is no left truncation, specify c_0=0. The values c_1 ,c_2 ,..., c_u are the right end points for the u intervals. The value c_u is the right-most point and is used as the WIDTH which defines the right truncation point. If all of the distances are less than or equal to c_u, the MCDS engine will not truncate data on the right unless RTRUNCATE is set. Perpendicular distance intervals can also be created for analysis with the NCLASS and WIDTH commands. NCLASS intervals of equal length are created between "Left" and "Width", if both NCLASS and WIDTH are given.

The SMEAR switch is used only if TYPE=LINE and radial distance/angle measurements were entered (DISTANCE=RADIAL). "Angle" defines the angle sector around the angle measurement and "Pdist" defines the proportional sector of distance to use as the basis for the smearing (see pg: 269-271 of Buckland et al. 2001).

If an observation is measured at angle "a" and radial distance "r", it is smeared uniformly in the sector defined by the angle range (a-angle,a+angle) and distance range (r*(1-pdist),r*(1+pdist)).

The NCLASS and WIDTH switches must also be given to define a set of equal perpendicular distance intervals. The proportion of the sector contained in each perpendicular distance interval is summed as an observation frequency and these non-integer frequencies ("grouped data") are analyzed to estimate detection probability.

Note: Distances specified by WIDTH, LEFT, and INTERVALS should be in the same units used for the entered data, even if the distance units are converted in the analysis.

Examples:

Truncate the distances at 100 feet, hence only use those less than or equal to 100 feet in the analysis. This value would be used even if the distances were converted to meters for analysis. The conversion is applied to the input width of 100 feet.

DIST /WIDTH=100;

The distance data were entered ungrouped but they were actually collected in these intervals; alternatively, to mediate the effects of heaping, these intervals were chosen to analyze the data.

DIST /INT=0,10,20,30,40,50,60,70,80,90,100;

The above example could also be entered as:

DIST/NCLASS=10/WIDTH=100;

ENCOUNTER Command

Syntax:

ENCOUNTER by | SAMPLE
STRATUM
ALL |

Description:

This command explicitly specifies that encounter rate should be estimated and the resolution at which the estimate(s) should be made (by SAMPLE, by STRATUM, or ALL data). This command is only necessary if density is not being estimated.

Examples:

A user wishes to explore the variability in encounter rate by listing the encounter rate for each sample. The variance of the encounter rate for each sample is assumed to be Poisson because the sample is a single entity.

```
ESTIMATE;
  ENCOUNTER by SAMPLE;
END;
```

A user wishes to explore the variability in encounter rate by listing the encounter rate for each stratum. The variance of the encounter rate for each stratum is computed empirically for each stratum with more than one sample; otherwise, it is assumed to be Poisson.

```
ESTIMATE;
  ENCOUNTER by STRATUM;
END;
```

A user wishes to only see the average encounter rate and an estimate of its variance. The variance of the encounter rate is computed empirically if there is more than one sample; otherwise, it is assumed to be Poisson.

```
ESTIMATE;
  ENCOUNTER ALL;
END;
```

ESTIMATOR Command

Syntax:

$$\text{ESTIMATOR} \quad /\text{KEY} = \begin{vmatrix} \text{UNIFORM} \\ \text{HNORMAL} \\ \text{NEXPON} \\ \text{HAZARD} \end{vmatrix} \quad /\text{ADJUST} = \begin{vmatrix} \text{COSINE} \\ \text{POLY} \\ \text{HERMITE} \end{vmatrix}$$

$$/\text{SELECT} = \begin{vmatrix} \text{SPECIFY} \\ \text{SEQUENTIAL} \\ \text{FOREWARD} \\ \text{ALL} \end{vmatrix} \quad /\text{ORDER} = O(1), O(2), ..., 0(\text{nap})$$

$$/\text{NAP} = \text{nap} \quad /\text{START} = A(1), A(2), ..., A(\text{nkp} + \text{nap})$$

$$/\text{CRITERION} = \begin{vmatrix} \text{AIC} \\ \text{AICC} \\ \text{BIC} \\ \text{LR} \end{vmatrix} \quad /\text{COVARIATES} = \text{cov1}, \text{cov2}, ...$$

$$/\text{LOWER} = \text{val1}, \text{val2}, ... \quad /\text{UPPER} = \text{val1}, \text{val2}, ...$$

$$/\text{ADJSTD} = \begin{vmatrix} \text{W} \\ \text{SIGMA} \end{vmatrix}$$

Description:

The ESTIMATOR command specifies the type of model for detection probability (g(x)) to estimate f(0) or h(0). The KEY switch specifies the key function to be used and the ADJUST switch specifies the type of adjustment function. The SELECT switch specifies the type of adjustment term selection which overrides the default value specified by the SELECTION command in the OPTIONS procedure (see the discussion on adjustment term in the introduction to the Estimate section).

If SELECT=SPECIFY is chosen, you can specify the number of adjustment parameters, the order of the adjustment term and starting values for the parameters. The number of adjustment parameters is set with the NAP (Number of Adjustment Parameters). NAP must be less than or equal to MAXTERMS - 'nkp' (number of key parameters). The orders of the adjustment term(s) are specified with the /ORDER switch. Starting values (/START) for the key and adjustment parameters can be given if the optimization algorithm suggests there are problems in finding the maximum of the likelihood function. The first 'nkp' starting values in the list should be the values for the key parameters and the remaining are for the 'nap' adjustment parameters. One reason for using the SELECT and NAP switches is to specify that only the key function should be fitted to the data. An example is given below.

CRITERION specifies the manner in which the number of adjustment terms is chosen for SELECT=FORWARD and SEQUENTIAL. LR specifies that a likelihood ratio test be performed using the PVALUE specified in OPTIONS. AIC specifies using the Akaike's Information Criterion for adjustment term selection.

COVARIATES gives a list of covariates that enter the scale parameter of the detection function – see the Users Guide Chapter 9 section Introduction to MCDS Analysis for more on this. Note that the covariates must be declared in the list of FIELDS in the Data section, and that factor covariates need to be declared as such using the FACTOR command.

The distances passed into the adjustment term formulae are scaled. ADJSTD determines how they are scaled – either by W (the truncation width) or SIGMA (the evaluated value of the scale parameter for this covariate). For a discussion of the difference, see Scaling of Distances for Adjustment Terms in Chapter 9 of the Users Guide.

LOWER and UPPER enable you to set bounds on the key function parameters (note that you cannot currently set constraints on adjustment term parameters).

If either are missing, default bounds are used (these are reported in the results). A value of -99 indicates use the default bound for that parameter. See below for an example.

Multiple ESTIMATORs can be specified within an ESTIMATE procedure and the "best" model is selected or estimates are given for each model (see PICK command). Note that if covariates are used, the same covariates must be declared in all ESTIMATOR commands – automatic selection among covariates is not currently supported.

The only portion of the command required is the command name ESTIMATOR because all of the switches have default values.

Default Values:

KEY = HNORMAL
ADJUST = COSINE
SELECT = SEQUENTIAL (or value set in Options section)
CRITERION = LR (except if **SELECT=ALL**)

Examples:

Use the following to fit a model with a half-normal key function (by default) and Hermite polynomials for adjustment functions. DISTANCE fits all possible combinations of adjustment terms and uses AIC to choose the best set of adjustment terms:

ESTIMATOR /ADJ=HERM /SEL=ALL;

Use the following to fit a model that uses the uniform key function with simple polynomial adjustment functions:

ESTIMATOR /KEY=UNIFORM /ADJ=POLY;

Use the following to fit a model that uses the hazard key without adjustments:

ESTIMATOR /KEY=HAZARD /SELECT=SPECIFY /NAP=0;

Use the following to fit a 2-term cosine series with terms of order 1 and 3 and specify the parameter starting values (note: nkp=0 for a uniform key):

ESTIMATOR /KEY=UNIF /SELECT=SPECIFY /NAP=2 /ORDER=1,3 /START=0.3,0.05;

Use the following to fit a hazard key with one polynomial adjustment, and a specified lower bound on the second parameter of the key function but the default lower bound of the first parameter. Imagine in this case that there are two strata and we want a lower bound of 2 on the 2nd parameter in the first stratum, and a lower bound of 2.5 on the 2nd parameter in the second stratum. Not a realistic example perhaps, but illustrates that you have to specify constraints separately for each stratum, as these are treated as separate key function parameters to be estimated. Also illustrates that you ignore the adjustment term parameters when setting bounds.

ESTIMATOR /KEY=HAZARD /ADJ=POLY /SELECT=SPECIFY /NAP=1 /LOWER=-99,2.0,-99,2.5;

G0 Command

Syntax:

G0=value /SE=value /DF=value;

Description:

This command assigns a value to g(0) which is assumed to be 1 unless a value is assigned with this command. The SE and DF switches are used to specify a standard error for the estimate so that estimation uncertainty of g(0) can be incorporated into the analytical variance of density. G0 is just a special case of a

multiplier, so see the MULTIPLIER command for details of its use and the options.

Default:

G0=1.0/SE=0.0;

Example:

G0=0.85/SE=0.12;

GOF Command

Syntax:

GOF /INTERVALS = c_0, c_1, c_2, … , c_u;

or

GOF /NCLASS=nclass;

or

GOF;

Description:

This command is used to specify the distance intervals for plotting a scaled version of the histogram of distances against the function g(x) (and f(x)) and for the chi-square goodness of fit test .

If the data are entered and analyzed ungrouped (/EXACT), the first 2 forms can be used to define the intervals which are used for plotting the data and for the chi-square goodness of fit test.The first form specifies the intervals exactly and the second form provides a shortcut approach of specifying "nclass" equal intervals (Note: the syntax from previous versions of GOF=c_0,c_1...; will also work). You can enter up to 3 of these commands to specify different sets of intervals. If you do not specify this command and the data are analyzed as ungrouped, 3 sets of intervals are constructed with equally spaced cutpoints and the number of intervals being the $n^{0.5}$ and 2/3 $n^{0.5}$ and 3/2 $n^{0.5}$.

If the data are entered grouped or entered ungrouped and analyzed as grouped (DISTANCE/INTERVALS= used in ESTIMATE) then only the third form can be used to specify that the GOF statistics should be generated. It is not possible to specify goodness of fit intervals other than those used to analyze the data

Examples:

Data are ungrouped and 2 different sets of intervals are specified.

GOF /NCLASS=5;
GOF=0,5,10,20,30,40,50;

Data are grouped but GOF statistics are desired.

GOF;

MONOTONE Command

Syntax:

MONOTONE = WEAK or **STRICT** or **NONE**

Description:

The estimators are constrained by default to be strictly monotonically non-increasing (i.e., MONOTONE=STRICT; the detection curve is either flat or decreasing as distance increases from 0 to w). In some instances, depending on the tail of the distribution this can cause a poor fit at the origin(x=0). Two options exist: 1) truncate the observations in the tail, or 2) use the command MONOTONE=WEAK; or MONOTONE=NONE;. MONTONE=WEAK; will

only enforce a weak monotonicity constraint (i.e., f(0) >= f(x) for all distances x). This will allow the curve to go up and down as it fits the data but it will not let the curve dip down at the origin. In some instances this will allow the estimator to achieve a better fit at the origin which is the point of interest. Setting MONTONE=NONE; will allow the curve to take any possible form except that it must remain non-negative.

Monotonicity is achieved by constraining the function at a fixed set of points. In some circumstances it is possible that the curve can be non-monotone between the fixed points. Typically, this results from trying to over-fit the data with too many adjustments with a long-tailed distribution. Truncate the data rather than attempting to over-fit.

Note that MONOTONE = NONE is the only allowed option when there are covariates in the detection function.

Default:

MONOTONE = STRICT; (no covariates)
MONOTONE = NONE; (covariates)

MULTIPLIER Command

Syntax:

MULTIPLIER = value1 /LABEL='name' /SE=value2 /DF=value3;

Description:

This command specifies a multiplier for the density and/or abundance estimate. Some uses of multipliers are discussed in the Users Guide section Multipliers in CDS Analysis.

Density/abundance is multiplied by the value of value1. The analytic variance estimate takes into account the additional variance due to the multiplier, as specified by value2, by adding an additional term ot the delta method formula (equation 3.70 in Buckland *et al.* 2001). The degrees of freedom for confidence limits are affected if a non-zero value is specified for value3, because an extra term is added to the Satterthwaite formula (equation 3.75 in Buckland *et al.* 2001).

/LABEL='name' – "name" is the name given to the multiplier in the output file

/SE=value2 – value2 is the standard error of the multiplier – use 0 if the multiplier value is known with certainty

/DF=value3 – value3 is the degrees of freedom associated with the multiplier – use 0 for infinite degrees of freedom.

Note that if you want a multiplier to divide the density estimate, simply specify value1 as the inverse of the multiplier value. Value2 (the SE) is then the multiplier SE divided by the square of the multiplier value.

There is no maximum to the number of multiplier commands within the Estimate section.

PICK Command

Syntax:

PICK = AIC or **AICC** or **BIC**

Description:

If more than one ESTIMATOR command is given a choice must be made as to which model will be used for the final estimate. The command PICK=AIC; instructs the program to choose the model that minimizes Akaike's Information Criterion, AICC minimizes the small sample corrected version of AIC, and BIC minimizes the Bayesian Information Criterion. If no command is given

PICK=AIC; is assumed. (Note: the option PICK=NONE, which told the program not to choose a model and to present the results of each, is no longer supported.)

If the BOOTSTRAP; command is given, the bootstrap is performed and the estimator is chosen for each analysis. Thus, even though a single estimator is chosen for the point estimate, different estimators can be chosen for each bootstrap and the standard errors and interval estimates incorporate the uncertainty of the model selection process.

Default:

PICK=AIC;

PRINT Command (Estimate section)

Syntax:

PRINT /YES=option list /NO=option list;

Definition:

This command can further expand or limit the output from the estimate procedure beyond what is defined by the PRINT command in the Options section. The PRINT command in the Options section allows hierarchical control of the output and defines the default values for this print command and thus retains its functionality. However, this command can be used to define whether each component is printed to the output file by specifying it either in the /YES or /NO list. The following are the values of the option list:

Options list	
All	Used in place of listing all options
Estimate	Density estiamte table
Explain	Explanation of estimation options
Fxest	Function parameter estimates
Fxfit	Function fitting/model selection
Fxiterations	Iterations of MLE
Fxplot	Function histogram/plots
Fxtest	Goodness of fit tests
Qqplot	QQ plots and associated statistics
Sbarest	Estimates of E(S)
Sbarplot	Size-bias regression plot

Below are listed the default values of the print options as defined by the value set by PRINT=in OPTIONS. (Y=Yes and N=No)

OPTIONS PRINT=

command value	ESTIMATE	FXEST	FXPLOT	FXTEST	SBAREST	SBARPLOT	FXFIT	FXIT	QQPLOT
SUMMARY	N	N	N	N	N	N	N	N	N
RESULTS	Y	Y	Y	Y	Y	Y	N	N	Y
SELECT	Y	Y	Y	Y	Y	Y	Y	N	Y
ALL	Y	Y	Y	Y	Y	Y	Y	Y	Y

By default, /YES=EXPLAIN is set to provide a printed explanation of the estimation options and models chosen (Note: in Versions prior to 1.20, the default was NO and it was set to YES with the more limited PRINT OPTIONS command).

SIZE Command

Syntax:

SIZE by | SAMPLE / STRATUM / ALL |

Details:

This command explicitly specifies that expected cluster size should be estimated and the resolution at which the estimate(s) should be made (by SAMPLE, by STRATUM, or ALL data). This command is only necessary if density is not being estimated or to specify a level of resolution different from density. The level of resolution for estimating cluster size must be less than or equal to the level for estimating detection probability, if a size bias regression estimate is computed.

Example:

A user wishes to examine detection probability and expected cluster size but not density at this point:

```
ESTIMATE;
  ESTIMATOR/KEY=UNIF;
  DETECTION ALL;
  SIZE ALL;
END;
```

VARN Command

Syntax:

VARN = POISSON or **b** or **EMPIRICAL**

Description:

This command specifies the type of variance estimation technique for encounter rate. The value POISSON specifies that the distribution of n (number of observations) is Poisson . EMPIRICAL specifies that the variance should be calculated empirically from the replicate SAMPLEs (section 3.6.2 of Buckland *et al.* 2001). If only one SAMPLE is defined in the data, the POISSON assumption is used unless a value b is specified. If a value b is specified it is used as a multiplier such that var(n)= bn (e.g., Buckland *et al.* 2001, section 8.4.1). The Poisson assumption is equivalent to specifying b=1. The default for VARN is EMPIRICAL unless there is only one SAMPLE, in which case, the default is POISSON.

Default:

VARN=EMPIRICAL

MCDS Engine Required Data Format

The distance sampling data should be stored in a data file. The location of this file is specified with the INFILE command in the Data section. The supported 7format for the data file is a flat file – i.e., a file containing columns that correspond with input fields and one row for each observation (and also transects without observations). Historical versions of this analysis engine used a hierarchical data format, but this is no longer supported. Columns should be separated by tab characters, and the FIELDS command should be used to tell the MRDS engine which column is which. Other commands may be required – see the Data section for details.

An example data file is given below:

```
Stratum A      100     Line 1A 10      14      F
Stratum A      100     Line 1A 10      8       M
Stratum A      100     Line 1A 10      22      M
Stratum A      100     Line 2A 10.3    7       F
Stratum A      100     Line 2A 10.3    37      F
Stratum A      100     Line 2A 10.3    13      F
Stratum B      123     Line 1B 5.7
Stratum B      123     Line 2B 8.4     27      M
Stratum B      123     Line 2B 8.4     76      F
Stratum B      123     Line 2B 8.4     44      M
Stratum B      123     Line 2B 8.4     7       M
```

Example data file

The corresponding FIELDS command for these data is:

```
FIELDS=STR_LABEL, STR_AREA, SMP_LABEL, SMP_EFFORT, DISTANCE, Sex;
```

This tells Distance that the first column is the stratum label, the second is the stratum area, the third is the sample label, the forth the sample effort, the fifth is the distances and the last is a column called "Sex". This last column will be used as a factor covariate, so the DATA section also needs the command

```
FACTOR=Sex /LEVELS=2 /LABELS='F','M';
```

Notice that for Line 1B there is nothing in the distance column – this is because no animals were seen on that line.

An easy way to generate an example data file, to get a feel for the required format, is to set up an analysis using the Distance graphical interface, and then run the analysis in **Debug mode**. In this mode, the Distance interface generates a command file and data file, and stores them in the Windows temporary folder, but does not run the analysis. For more about Debug mode, see the Program Reference page on the Analysis Preferences Tab. This strategy will also enable you to see what commands are required in the Data section for a particular data file.

Output From the MCDS Engine

Output from the MCDS engine takes two forms:

- MCDS engine command line output. A number returned to the command line when the run finishes, giving the status of the run. Occasionally, other output, such as FORTRAN error or warning messages, may appear there as well.
- Up to 6 results files, as specified in the header section of the command file. These files are:
  - MCDS engine output file
  - MCDS engine log file
  - MCDS engine stats file
  - MCDS engine bootstrap file
  - MCDS engine plot file
  - MCDS engine bootstrap progress file

The format of each of these outputs is given in the following sections.

MCDS Engine Command Line Output

Run status

When the MCDS engine is run from the command line, it returns a number when the run finishes. This number gives the status of the run, as follows:

- 1 means the analysis ran OK
- 2 means it ran with warnings (see log file for details)
- 3 means it ran with errors (see log file for details)
- 4 means it ran with file errors (e.g., could not find the specified command file)
- some other number. A major error occurred (see below).

These numbers are also returned if the engine is run from another program as an independent process, and so can be used by the program to diagnose whether the run was OK.

A number returned to the command line when the run finishes, giving the status of the run. Occasionally, other output, such as FORTRAN error or warning messages, may appear there as well.

FORTRAN Debugging output

Occasionally, some other text is written to the standard output, which is usually the command line, by the FORTRAN run time library used to run mcds.exe. A mild example is that the number of underflow errors are written out, e.g.,:

```
forrtl: error (74): floating underflow
forrtl: error (74): floating underflow
forrtl: info (300): 30 floating underflow traps
```

Floating point underflow occurs when a number is calculated that is smaller than the smallest number the computer can store and the number is instead stored as zero. This rarely causes problems in practice – although it is worth double-checking your results.

A more extreme example is if there is a program crash, debugging information is written out. If this happens, a copy of the Distance project or command file should be sent to the program authors. In the example below, the program crashed with a "floating invalid" error on line 398 of the routine SBREG, which was called from line 205 of CMOD, etc.

```
forrtl: error (65): floating invalid
Image         PC        Routine         Line   Source
MCDS.exe      0040DA68  SBREG             398  Cmod.for
MCDS.exe      0040C294  CMOD              205  Cmod.for
MCDS.exe      0040B2F3  ESMOD              29  Cmod.for
MCDS.exe      0044763C  ESTPARM           487  Estmte.for
MCDS.exe      00446471  ESTMTE            291  Estmte.for
MCDS.exe      00445201  ESTPROC            88  Estmte.for
MCDS.exe      004136EA  CNTRL             113  Control.for
MCDS.exe      004381CF  DISTANCE          263  Distance.for
MCDS.exe      004F2459  Unknown       Unknown  Unknown
MCDS.exe      004C8BE3  Unknown       Unknown  Unknown
kernel32.dll  7C816FD7  Unknown       Unknown  Unknown
```

MCDS Engine Output File

This file is used in the Distance interface to fill the Results tab of the Analysis Details Window. It is divided into a number of pages, and each page title is delimited by tab characters:

[Tab]Page title[Tab]

This file is designed for human reading, not machine processing (except that pages can easily be separated by searching for two tabs with text in between). Most important statistics are output in a compact, easily parsed format in the

MCDS Engine Stats File, which should be the first port of call for extracting results by machine.

MCDS Engine Log File

This file is used in the Distance interface to fill the Log tab of the Analysis Details Window. It contains a copy of the input commands, together with output from the analysis engine about any errors or warnings encountered while processing the commands. This file is the first place to look if the analysis engine returns a warning or error value (2 or 3) after a run.

A comprehensive list of the warning and error messages is given in the section MCDS Engine Error and Warning Messages.

MCDS Engine Stats File

This file contains a compact output of summary statistics. The Distance interface uses this file to extract data for the Analysis Browser table. A set of records is output for each model defined in the Detection Function Models. Each record is given a new line. The record structure is as follows:

Stratum	stratum number (or 0 if the estimate is for a sample or all data).
Sample	sample number (or 0 if the estimate is for a stratum or all data).
Estimator	number of the estimator (in the order given in the Estimate procedure)
Module	number of the parameter module (see below)
Statistic	number of the statistic within the parameter module (see below)
Value	estimate value
Cv	coefficient of variation of estimate or 0.0
Lcl	lower confidence limit or 0.0
Ucl	upper confidence limit or 0.0
Df	degrees of freedom for interval or 0

The modules and statistics within each module are listed in below in the order in which they are summarized in the output. The FORTRAN format for each record is:

```
FORMAT(2(1X,I5),2(1X,I1),1X,I3,5(1X,G14.7))
```

This is different from the format for previous versions of Distance

Each field is separated by a space, so the records can be read into a spreadsheet or other program as space delimited or as fixed-width format. The record for a module/statistic type is only output if it is relevant and it was computed in the analysis.

The following table defines the module and statistic codes used:

Module	Statistic/Parameter Estimate
1 – encounter rate	1 – number of observations (n) 2 – number of samples (k) 3 – effort (L or K or T) 4 – encounter rate (n/L or n/K or n/T)

	5 – left truncation distance 6 – right truncation distance (w)
2 – detection probability	1 – total number of parameters (m) 2 – AIC value 3 – chi-square test probability 4 – f(0) or h(0) [1] 5 – probability of detection (P_w) [1] 6 – effective strip width (ESW) or effective detection radius (EDR) [1] 7 – AICc 8 – BIC 9 – Log likelihood 10 – Kolmogorov-Smirnov test probability 11 – Cramér-von Mises (uniform weighting) test probability 12 – Cramér-von Mises (cosine weighting) test probability 13 – key function type [2] 14 – adjustment series type [3] 15 – number of key function parameters (NKP) 16 – number of adjustment term parameters (NAP) 17 – number of covariate parameters (NCP) 101 … (100+m) – estimated value of each parameter [5]
3 – cluster size	1 – average cluster size [1] 2 – size-bias regression correlation (r) 3 – p-value for correlation significance (r-p) 4 – estimate of expected cluster size corrected for size bias [1]
4 – density/abudance	1 – density of clusters (or animal density if non-clustered) [1] 2 – density of animals [1] 3 – number of animals, if survey area is specified [1] 4 – bootstrap density of clusters [1, 4] 5 – bootstrap density of animals [1, 4] 6 – bootstrap number of animals [1, 4]

[1] Values for CV, LCL, UCL and DF are included for these statistics.
[2] Key function types are: 1 = uniform, 2 = half-normal, 3 = negative exponential, 4 = hazard rate
[3] Adjustment series types are: 1 = simple polynomial, 2 = Hermite polynomial, 3 = cosine
[4] Bootstrap CV calculated as bootstrap SE / bootstrap point estimate; df field here is the number of bootstraps
[5] Statistic 101 corresponds with the parameter identified as A(1) in the results, 102 with A(2), etc.

MCDS Engine Plot File

This file contains the data used to construct the high resolution qq and histogram plots in the Distance interface. The output format is:

```
Title line - plot 1 (up to 80 char)
Sub-title line (up to 60 char)
x-label (up to 30 char)
y-label (up to 30 char)
# of data rows (r)
x1,y11,y21,y31,y41
x2,y21,y22,y31,y41
.
.
.
xr,yr1,yr2,y31,y41
Title line - plot 2 (up to 80 char)
Sub-title line (up to 60 char)
x-label (up to 30 char)
y-label (up to 30 char)
..
etc
```

The number of columns of data (yr1, yr2, etc) depends on the plot:

- For qq-plots there are 4 columns: 1 and 2 are the x and y coordinates of the data points (i.e. the edf and the fitted cdf); 3 and 4 are the x and y coordinates of the line that runs from (0,0) to (1,1). If you are using this file to recreate the plot in another package, you could easily ignore columns 3 and 4 and replace them with a (0,0) (1,1) line.
- For plots containing the data histograms and accompanying pdf or detection function plots there are 4 columns: 1 and 2 give the x and y coordinates which, when joined up, give the fitted detection function or pdf and 3 and 4 give a set of x and y coordinates which, when joined up, produce the data histograms
- For the MCDS example detection function plots, which contain 3 detection functions, there are 6 columns: 1 and 2 give the x and y coordinates for the first detection function, 3 and 4 give this for the second detection function and 5 and 6 give the coordinates of the third detection function.

You can see an example of these kind of data being used to produce a plot in a tip under Exporting CDS Results from Analysis Details Results in Chapter 8 (although the data there come from copying the plot to the clipboard rather than directly from the plot file).

MCDS Engine Bootstrap File

This file has the same format as the MCDS Engine Stats File, but contains one set of stats records for each bootstrap. These records are appended one after another as the bootstrap progresses.

MCDS Engine Bootstrap Progress File

This file is intended to be read periodically during a bootstrap analysis. It is only created if a bootstrap is requested, and if the file name is not "None". If the file exists already, it is overwritten.

The file is empty until the bootstrapping starts. Then, it contains a 3-digit integer (between 0-100) which indicates the percentage of the way through the bootstrap we are.

An example of the use of this file is the Distance interface, which reads it every second during a bootstrap analysis and uses the contents to add a "Progress x%" message after the "Running analysis x" on the status line of the main toolbar.

MCDS Engine Limitations

The following limitations apply to the MCDS engine:

Limitation	Maximum number
Observations	100 000
Samples (transects)	50 000
Strata	1 000
Cutpoints in GOF and interval data	25
Detection function models	5
Adjustment terms per model	5
Covariates	10
Levels, for factor covariates	200

In addition, constraints cannot be set on the detection function if there are covariates in addition to Distance (see next section).

MCDS Engine Fitting Algorithms

CDS Analyses

For analyses with distance as the only covariate in the detection function (CDS analyses), the detection function is fit using the constrained maximum likelihood method outlined in section 3.3.5 of Buckland et al. (2001). The fitting algorithm used is subroutine DNCONG, from the IMSL FORTRAN 90 Mathematics and Statistics Library version 3.0, written by Visual Numerics Ltd. This subroutine uses a successive quadratic programming algorithm to minimize the negative log-likelihood subject to monotonicity constraints on the detection function.

MCDS Analyses

For fitting the detection function when there are one or more covariate, the MCDS engine uses an iterative maximum likelihood routine, based on the Newton-Raphson algorithm. It alternates between fitting the key and covariate parameters conditional on the current estimates for the adjustment parameters, and fitting the adjustment parameters conditional on the current estimates for the key and covariate parameters. When close to convergence, it switches to maximizing all parameters at once.

This algorithm cannot cope with constraints, and therefore no monotonicity constraints can be enforced with the MCDS engine.

In some cases, the user specifies the MCDS engine but there are no covariates. For example, imagine that the detection function is to be fit by stratum, and that there is one factor covariate with two levels. In one stratum, only one level of the factor occurs in the data. In this stratum, no covariate parameters. When this occurs, the CDS algorithm is used to fit the data, but still with no monotonicity constraints.

This means that if you select the MCDS engine in the Distance interface, but do not specify any covariates, you will get identical output to selecting the CDS engine with no monotonicity constraints.

MCDS Engine Error and Warning Messages

This section gives a comprehensive list of the warning and error messages that can be generated by the MCDS engine, and can occur in the output or log file. Explanations for some of the messages are given – please contact us if you need an explanation for a message we don't explain here, or get a message that is not documented (it's possible we missed some!).

The standard format for an error or warning message is
`** [Bootstrap] level: message **`
where `level` can be either "Warning", "Error" or "Internal Error", `message` is the text of the message, and the word `Bootstrap` appears if the problem occurred while running a bootstrap replicate dataset.

Some error and warning messages have been discussed on the distance-sampling email list, so it's worth searching the list archives to for more information, including possible remedies.

MCDS Engine Warning Messages

These are used to indicate a mild problem with the analysis (such as small sample size, possible lack of convergence, etc.), or to alert the user to an unusual aspect of the analysis that they should be aware of. Warnings do not result in an analysis terminating, but should be taken seriously as the quality of the results may be compromised.

Many of these messages will not be seen by users of the Distance interface, as they relate to mutually incompatible combinations of commands that are already screened for by the interface.

ID	Warning Message	Explanation
1	Angle not valid for TYPE=POINT	Mutually incompatible commands
2	SIZE not valid for OBJECT=SINGLE	Mutually incompatible commands
3	The previous stratum had no samples so it will be ignored.	
4	An estimator was not chosen. The default estimator will be used.	
5	Cannot estimate encounter rate variance empirically when estimating encounter rate by sample. Sample level encounter rate variances will be estimated assuming distribution of observations is Poisson.	
6	Degrees of freedom less than 1.0 for estimating confidence limits on density of individuals. Confidence limits not calculated.	
7	Degrees of freedom less than 1.0 for estimating confidence limits on F0. Confidence limits not calculated.	
8	Degrees of freedom less than 1.0 for estimating confidence limits on density of objects. Confidence limits not calculated.	
9	Warning: Estimated correlation between	

	parameters > +/- 1.0.	
10	Estimation routine failed to converge due to negative area estimates on iteration [iteration number]. Using results from previous iteration.	
11	Estimation routine failed to converge due to singular information matrix on iteration [iteration number]. Using results from previous iteration.	
12	Estimation routine failed to converge.	
13	FIELD will be ignored in the data	
14	INTERVALS switch ignored because NCLASS specified.	Mutually incompatible commands
15	Missing item in list two adjacent commas [value here]. Skipping to next item.	
16	Negative variance estimate for f0. Invalid variance. Results may not be reliable.	
17	One or more cluster sizes are 0. These observations will be used in cluster size estimation. If you intended to code these values as missing, please enter them as -1.0	See Zero Cluster Sizes in CDS Analysis in the Users Guide.
18	One or more cluster sizes is coded as -1. Distance assumes -1 to mean a cluster of undetermined size. These observations are used for estimating detection probability and encounter rate, but not cluster size.	See Missing Cluster Size Data in CDS Analysis in the Users Guide.
19	Parameter [parameter number] is at a lower bound.	
20	Parameter [parameter number] is at an upper bound	
21	Parameters are being constrained to maintain a positive pdf	
22	Parameters are being constrained to obtain monotonicity.	
23	Previously read samples were not assigned a stratum, so all strata will be ignored.	
24	SIZEC is an invalid command when OBJECT=SINGLE, and so was ignored.	Mutually incompatible commands
25	Some of the estimates of f0 are negative. Results are not reliable.	
26	Some parameters are very highly correlated.	
27	The /BIAS switch is not allowed when cluster size is a covariate, and so it has been ignored.	Mutually incompatible commands
28	The cluster size covariate is a factor and so it is assumed that factor levels correspond to cluster sizes.	
29	The estimated analytic variance for f0 gives a CV greater than 10000%, and hence results may not be reliable. To avoid numerical problems, the CV for f0 was assigned a value of 9999.99%.	
30	The estimated area is negative and only a single iteration of the estimation routine has been carried out.	
31	The number of lower bounds for the parameters does not match the number of parameters in the model. The default bound for parameter A [parameter number] is being used instead.	
32	The number of starting values for the parameters	

	does not match the number of parameters in the model. The default starting value for parameter A [parameter number] is being used instead.	
33	The number of upper bounds for the parameters does not match the number of parameters in the model. The default bound for parameter A [parameter number] is being used instead.	
34	The starting value for parameter A [parameter number] must be > 0. Using the default starting value.	
35	There are less than 10 data points per estimated parameter. Results may not be reliable.	
36	There is only one level for factor covariate [covariate]. A minimum of two levels is required for estimation; hence this covariate will be omitted.	
37	There is only one level for factor covariate [covariate]. A minimum of two levels is required for estimation; hence this covariate will be omitted from estimates for stratum [stratum] sample [sample].	
38	There is only one level for factor covariate [covaraite]. A minimum of two levels is required for estimation; hence this covariate will be omitted from estimates for stratum [stratum]	
39	When cluster size is a covariate, variance of the cluster size, density of individuals, and abundance estimates can only be obtained via the bootstrap. You have not specified the bootstrap variance option, so these variance estimates will not be produced.	See Cluster Size as a Covariate in the Users Guide.
40	When cluster size is a covariate encounter rates are not computed.	See Cluster Size as a Covariate in the Users Guide.
41	When cluster size is a covariate no stratification is allowed.	
42	When covariates are being used, a number of intervals > 20 for GOF tests may cause the program to terminate with an error.	
43	The number of intervals - 1 = [number] which is less than the number of key parameters [number]. No fit possible.	
44	The number of intervals - 1 equals the number of key parameters [number] so no adjustments can be made.	
45	The number of observations - 1 equals the number of key parameters [number], so no adjustments can be made.	
46	There are no cluster size observations selected. Cannot estimate expected cluster size.	
47	Negative variance for expected cluster size. No size bias adjustment. Average cluster size used instead.	
48	Convergence failure.	
49	Estimated cluster size greater than exp14. Average cluster size used instead.	
50	SEED cannot be a negative number. It has been set to 0	

51	Negative variance estimate for parameter. Invalid variance.	
52	Number of cluster size measurements = [number]. This is not sufficient for size-bias regression. Average cluster size used instead.	
53	Number of cluster size measurements = [number] This is not sufficient to estimate a mean and variance	
54	Number of observations is small. Do not expect reasonable results.	
55	Size bias adjustment has increased expected cluster size.	
56	The number of adjustment parameters allowed has been reduced to [number] because of limited number of observations.	
57	Too few observations to calculate AICc. AICc set to 0.	
58	Too few observations. An estimate of f0 cannot be computed. f0 set to 1/width.	
59	Two models have the same [model selection statistic]. Choosing one of them at random.	
60	Zero observations. An estimate of f0 cannot be computed. f0 set to 0.	
61	Angle not valid for DIST=PERP	Mutually incompatible commands
62	Area = 0 for stratum	
63	BOOTSTRAPS may not exceed 5000. Set to 5000	
64	BOOTSTRAPS should be at least 100	
65	Cannot specify CONVERT without MEASURE and UNITS',/, CONVERT value will not be used.	Mutually incompatible commands
66	CONVERT value will overide previous value',' specified.	
67	INTERVALS ignored because NCLASS was set.	Mutually incompatible commands
68	Warning: Invalid or missing covariate.	
69	NCLASS and INTERVALS both set. INTERVALS',' ignored.	Mutually incompatible commands
70	No observations in stratum [stratum] so estimating f0 using global average f0|z. Results are therefore not reliable.	
71	SEED should be an odd number greater than 2000000	
72	Warning: Seed will be set with value from clock	Refers to the random number seed. This warning does not occur when calling MCDS from the interface as SEED=0 is specified, which means set from clock. This warning only occurs if SEED is not specified, and is intended to remind the user the seed has come from the clock.
73	SMEAR switch only valid for ungrouped dist/angle measurements.	Mutually incompatible commands

74	SMP_EFFORT not in data. Assumed to be 1 for each point	
75	Specified width [width] does not match an interval value. It has been set to [new width]	When truncating data in intervals.
76	There is only one level for factor covariate [covariate]. A minimum of two levels is required for estimation; hence this covariatewill be omitted from estimates for sample [sample]	
77	Warning: TITLE value not found.	
78	Too many sets of GOF, only [number] allowed.	Currently, [number]=3
79	User is overriding a conversion factor available in the program.	Obsolete – conversion factors between units now always specified by Distance interface, so this warning is suppressed.
80	For goodness-of-fit interval set [number]. Number of goodness-of-fit intervals reduced.	
81	For goodness-of-fit interval set [number]. Interval end-point modified to match width.	
82	For goodness-of-fit interval set [number]. Goodness-of-fit intervals specify testing subset of the data.	
83	For goodness-of-fit interval set [number]. Specified intervals are inconsistent with width.	
84	For goodness-of-fit interval set [number]. Interval begin-point modified to match left truncation.	
85	For goodness-of-fit interval set [number]. Specified intervals are inconsistent with left truncation value.	
86	Exact distance values, rather than distance intervals have been used in size bias regression calculations.	
87	No monotocity constraints are allowed when covariates are present.	Mutually incompatible commands

MCDS Engine Error Messages

These are used to indicate a severe problem with the analysis – enough to make the results invalid. These are sometimes usually enough to terminate the entire analysis, but more usually result in a part of the analysis being terminated (for example a row of data is dropped or a detection function estimate is not used).

Many of these messages will not be seen by users of the Distance interface, as they relate to mutually incompatible combinations of commands that are already screened for by the interface.

ID	Error Message	Explanation
1	Total of cluster frequencies = [number] Exceeds maximum number of observations – [number] Procedure terminated.	
2	Number of adjustment parameters NAP is greater than MAXFIT.	Too many adjustment terms specified.
3	Number of parameter bounds exceeds [number]	More parameter bounds than parameters.
4	Ambiguous value for command [command]	
5	Invalid value. Use either Yes, No, On, Off, True,	

	False.	
6	Invalid value for command [command]	
7	[Command] is an invalid command for HIERARCHY structure.	Obsolete.
8	A maximum of [number] levels is allowed for each factor covariate.	See MCDS Engine Limitations.
9	A maximum of 10 covariates may be specified.	
10	ANGLE needed in data	
11	Area under fx or gx is zero.	
12	At most one group contains observations.	
13	Bootstrap will not be done because', observations are not being re-sampled and density estimated by sample	
14	Cannot scale distances by sigma when using the uniform key function: change /ADJSTD option to W.	
15	Cannot use multiple DETECTION commands for CDS analysis or when cluster size is a covariate.	
16	Cluster size frequency < 0 [value of cluster freq]	
17	CONFIDENCE must be between 1 and 99	
18	Covariate specified in the ESTIMATE command but not in the DATA command: [covariate],	
19	CUERATE must be a positive number.	
20	Dataset has been cleared.No data has been stored.	
21	Density for each sample is unnecessary when detection and expected cluster size are estimated at higher levels	
22	Detection probability must be estimated for size bias calculations	
23	Distance frequency < 0 [value of distance freq]	
24	DISTANCE needed in data	
25	Due to errors, this ESTIMATOR command will be ignored.	
26	Due to errors, this GOF command will be ignored.	
27	Error reading in values.	
28	Exceeded array size [size] - for entering data.	
29	Exceeded maximum number of cluster size observations = [number]	Same as number of distance observations, below.
30	Exceeded maximum number of distance', observations = [number]	See MCDS Engine Limitations.
31	FIELDS have not been set. Data cannot be read.	
32	Filename on INFILE command was not found.	
33	FLAT data structure invalid for grouped data	
34	Incompatible resolution levels for estimation of [one component of estimation] and [another]	
35	Incorrect number of data values.	
36	Interval values for cluster sizes are out of order.	Obsolete
37	Interval values for distances are out of order.	Obsolete
38	Intervals for clusters cannot be specified because	Obsolete

	the data were entered in intervals.	
39	Intervals for distance cannot be specified because the', data were entered in intervals.	Obsolete
40	Invalid cluster size < 0 [value]	
41	Invalid command encountered – [command]	
42	Invalid distance <0 [value]	
43	Invalid filename or file could not be found [filename]	
44	Invalid initial values for parameters	
45	Invalid option for CUERATE command.	
46	Invalid or missing angle.	
47	Invalid or missing cluster size.	
48	Invalid or missing distance.	
49	Invalid or missing value for [variable]	
50	Invalid or missing value for sample effort = [sample name] Sample will be ignored.	
51	Invalid radial distance <0 [value]	
52	Invalid smearing angle = [value] It must be > 0 & <90	
53	Invalid smearing distance. It must be > 0.	
54	Invalid value for adjustment ORDER = [value]	
55	Invalid value for NCLASS. It must be between 2 and [max number of classes]	See MCDS Engine Limitations.
56	Invalid value for sighting angle <0 OR >360 = [value]	
57	ITERATIONS must be > or = to 25.	
58	Left truncation value cannot be negative.	
59	LENGTH command is invalid for point transect data.	Mutually incompatible commands
60	Maximum number of samples [number] exceeded.	See MCDS Engine Limitations.
61	Maximum number of strata [number] exceeded.	See MCDS Engine Limitations.
62	Mismatched number of observations for multiple measurements.	
63	Missing sample label. Further data will be ignored.	
64	More cluster size frequencies were given than intervals	
65	More distance frequencies were given than intervals.	
66	More than 10 multipliers were specified. Excess will be ignored.	
67	Multiplier value must be > 0	
68	NCLASS and WIDTH setting needed for SMEAR switch.	
69	NCLASS must be > 1 and <= [max value]	See MCDS Engine Limitations.
70	NCLASS set without a WIDTH value, both must be set.	Mutually incompatible commands
71	Negative variance estimate for f0. Invalid	

	variance.	
72	No data available to be analyzed.	
73	Not a valid option with grouped data – [option]	Mutually incompatible commands
74	Number of adjustment parameters NAP is greater than maximum possible for this model = [max. number possible for this model]	
75	Number of starting values exceeds [number of parameters]	
76	One or more estimated CDF is less than zero. Value has been set to zero.	
77	Only a single covariate may be specified as the cluster size covariate.	
78	Re-sampling unit Strata, Sample, Obs not set. Bootstrap will not be attempted.	
79	SE for multiplier must be non-negative number.	
80	SF must be between 0.0 and 1.0	
81	SMP_LABEL needed in data	
82	Specified variance for N must be >= 0.Value entered was [value]	
83	Standard Error for CUERATE must be a positive number.	
84	Strata can not be re-sampled if Density by Strata	
85	The [command] command is only valid when there are covariates.	
86	The FIELDS command must be specified before the [command] command.	
87	The number of adjustments specified by ORDER [value] does not match the number specified by NAP [value]	Mutually incompatible commands
88	The number of adjustments specified by ORDER exceeds the maximum of [maximum]	
89	The number of covariates which are factors cannot exceed the total number of covariates to be included.	
90	The significance level for the test must be in the interval 0-1.	
91	The specified factor covariate must match one of the covariates given in the FIELDS', command.	
92	The total number of cluster sizes does not match the total number of distances.	
93	The truncation proportion must be >=0 and <1.	
94	The width must be a positive value.	
95	There must be 2 parameters for smearing. 'Smear=angle,dist'	
96	There were [value] covariates specified in the', ESTIMATE command, but only [value] specified in the DATA command. The latter must be >= the former.	
97	Unexpected end-of-file encountered.	
98	Unknown units for distance measurement. Need to set conversion factor or correct input.	
99	Unknown units for distance/area conversion. Need to set area conversion factor or correct	

	input.	
100	Unknown units for distance/length/area conversion. Need to set conversion factor or correct input.	
101	Valid values for LOOKAHEAD are 1 – [max lookahead]	
102	Values for lower parameter bounds must be smaller than values for upper parameter bounds!	
103	When cluster size is a covariate no stratification is allowed.	
104	When covariates are present, only the half-normal or hazard-rate keys may be specified.	
105	When using the FACTOR command, the switches /LEVELS, and /LABELS must be specified.	
106	WIDTH must be given with NCLASS value.	
107	You have requested more estimators than the maximum of [maximum estimators]	
108	You must reset FIELDS or OPTIONS	
109	PVALUE must be between 0.0 and 1.0	
110	EPS must be BETWEEN 0.1 and 1.0E-8.	
111	Exceeded maximum array storage	
112	Maximum number of observations – [max obs] exceeded. Procedure terminated.	
113	Negative variance estimate for f0. Invalid variance.	
114	Negative variance estimate for parameter. Invalid variance.	
115	Negative variance estimate for part of f0|Z. Invalid variance.	
116	Bad bootstrap sample.	

MCDS Engine Internal Error Messages

These occur when the MCDS engine encounters a situation that should not occur under normal circumstances – for example a negative estimated effective strip width. The analysis is aborted and no useful results are produced. If you encounter such an error, please contact the program authors, sending a copy of your command and data files, or Distance project.

ID	Internal Error Message	Explanation
21	invalid ordering sign.	
22	CN = [value] INC = [value]	
23	ier – [list of values]	
24a	G0 = 0, MU = [value]	
24b	MU = 0, G0 = [value]	
24c	WIDTH=0.0	
25	fx not found [list of values]	
26	Normalizing factor MU2=0.	
27	AREA divisor for plot=0.	
28	N equal 0 for plot.	
29	Plot interval length for x = 0.	

30a	Problems with incomplete gamma fct (GSER)	
30b	Problems with incomplete gamma fct (GSF)	
30c	Problems with incomplete gamma fct (GAMMP) x or a <=0	
31	Invalid degrees of freedom or chi-square value.	
32	variance of cluster sizes is <= 0	
33	Normalizing factor MU1=0.	
34	Area under PDF=0 for at least one observation.	
35	Confidence limit ZVAL is NaN. Confidence limits set to 0. DF= [value]	
36	F0NOBS= [value]	
37	N= [value]	
38	WIDTH has been erroneously set to zero	
39a	Standard error is 0 for one of the parameters.	
39b	Standard error is 0 for parameter [parameter]	
39c	Cannot scale distances by sigma for uniform key function	
40	DERIVS - PJ = 0 value	
41	DERIVS - MU = 0	
42	PROB - MU = 0	
43	Invalid N in inverse routine = [value]	
44	WIDTH is zero.	
45	Attempt to divide by zero in inversion.	
46	compute pdist error [values]	
47	compute pdist [values]	
50	Mismatched strata – [values]	
51	Mismatched samples – [values]	
53	Mismatched modules – [values]	
54	Mismatched stats – [values]	
61	F0 = 0.0	
62	N= [value]	
70	CLEVMULT=0	
71	COVLEVELSI=0	
91	Problems with settings for estimation routine.	
102	Could not evaluate area under CDF.	

Changes in MCDS Engine Since Distance 2.2

- Interactive mode not supported – use only in batch mode
- Flat file data entry added, and the old nested style is no longer supported. The DATA statement should always have the option /STRUCTURE = FLAT
- Changes in commands:
  - New commands in DATA section: FIELDS, FACTOR and SIZEC.

- New switch /COVARIATES in ESTIMATOR command. Note that covariates must be the same in all ESTIMATORS in the same run.
- ASSIGN commands not supported – assign output files through the first 6 lines of the command file – see Header Section.
- HELP commands no longer supported
- SQUEEZE command no longer supported
- New MULTIPLIER command
- DF added to the CUERATE command (which is, after all, just another multiplier).
- Not recommended to use the SF command – use MULTIPLIER instead
- Model fitting commands EPSILON and ITERATIONS removed
- LIST command in Data section no longer supported
- In GOF /SAS and /SPLUS switches no longer supported
- VARF command no longer supported
- PICK=NONE no longer supported.
- When bootstrapping, point estimate is mean of the bootstrap replicate point estimates.

- Changes in output format:
  - Output file pages are no longer separated by page break characters
  - Page titles in the output file are surrounded with tab characters to enable them to be easily recognized by a regular expression parser
  - Format of the stats and bootstrap stats file changed (each line is longer) – see MCDS Engine Stats File.
  - Extra output in the stats file – e.g., parameter estimates
  - Bootstrap progress file added to give a way to allow the user to find out how far the bootstrap has progressed.

Appendix – HT estimation of density when probability of coverage is unequal

Review of the Horvitz-Thompson-like estimator

This appendix, and the associated analysis engine is new in Distance 6.0 Beta 5. This text will eventually be moved to become a chapter in the Users Manual.

This entire chapter is for advanced users only!

The theory behind this estimator of animal abundance is described in Strindberg et al. (2004) and much of the terminology associated with this method can be found in the Glossary of Terms at the end of this users guide.

The DHT engine currently has the following capabilities. We expect to extend these in future versions.

- Line transect surveys (has not been tested for strip or point transects)
- Coverage probability for each detected object is determined by the survey design engine in Distance
- Estimation of density/abundance using a detection function fitted in a previous analysis – allows different subsets of the data to be used for encounter rate and detection function fitting (see Using a Previously Fitted Detection Function to Estimate Density in MRDS).

The basic form of the Horvitz-Thompson-like estimator employed by the DHT engine is

$$\hat{N} = \sum_{i=1}^{n} \frac{s_i}{\hat{p}_i \hat{q}_i}$$

where s_i is the size of the i^{th} cluster, $\hat{p}_i$ is the estimated probability of detection of the i^{th} cluster, and $\hat{q}_i$ is the estimated coverage probability of the location where the i^{th} cluster was located. This formulation is premised on the basis of estimating abundance of individuals; if interest was instead focused upon the

estimation of abundance of clusters in the population, a "1" would be substituted for the numerator.

The job of the DHT engine is to bring together these three pieces of information to create an estimate of abundance. In its simplest incarnation, the cluster size is observed by the researcher, coverage probability is estimated from the survey design (by the survey design engine in Distance prior to the conduct of the survey), and finally the detection probability is estimated by fitting a detection function to the observed distances.

As with the MRDS and the DSM engine, the DHT engine is implemented as a library in the free statistical software R. When you run a DHT analysis from Distance, Distance creates a sequence of R commands, calls the R software, waits for the results and then reads them back in. Therefore, before you can use the DHT engine, you must first ensure that you have R correctly installed and configured. For more on this, see R Statistical Software in Chapter 7 of the Users Guide.

To produce an abundance estimate based upon unequal coverage probability in Distance, you then need to set up the project appropriately and include data in the correct format – see Setting up a Project for DSM Analysis. You must next create one or more model definitions using the MRDS analysis engine, and associate these model definitions with analyses to derive detection probabilities for each objected detected. For more about the basics of setting up analyses, see Chapter 7 - Analysis in Distance. More details of the various models available in the MRDS engine are given in Defining MRDS Models, and a detailed description of the options available in the Model Definition Properties pages for this engine is given in the Program Reference pages Model Definition Properties - MRDS. After deriving detection probabilities, coverage probabilities associated with each detected object is derived from the coverage grid constructed by the survey design engine.

If you are new to Distance, we strongly recommend you familiarize yourself with the CDS analysis engine, for example by working through Chapter 3 - Getting Started before trying to analyze DHT data.

The Williams multi-species survey sample project is an example of how to set up a double-observer study and specify analyses – see Sample Projects.

In this chapter we also provide some analysis guidelines, give a list of the output the engine can produce and cover various miscellaneous topics.

If you are familiar with the R software, you can run the DHT engine directly from within R, bypassing the Distance interface altogether. For more information, see Running the DSM Analysis Engine from Outside Distance.

Setting up a Project for DHT Analysis

The premise of this documentation is that users have employed the survey design in Distance (see Chapter 6 - Survey Design in Distance) to lay out a survey and then that survey has been carried out in the field. In the process of designing the survey, the user generated a coverage grid that is stored as a layer within the Distance project. Having performed the survey, information is entered into the

Observation layer of the Distance project in the usual fashion, with the additional data requirement that the location of each detection (x and y coordinates) is also recorded in the Observation layer.

The easiest way to set up a new project for a DHT analysis is using the Setup Project Wizard.

- In Step 1, under I want to: select Analyze a survey that has been completed.
- Be sure to tick the box indicating the Project will contain geographic information at the bottom of the Step 1 screen
- In Step 3, under Observer configuration, select Double observer. But see also Single Observer Configuration in the MRDS Engine.
- Follow through the rest of the wizard as usual.

Distance then creates the appropriate data fields for double observer data, and you can then import your data using the Import Data Wizard.

Alternatively, you can create the appropriate fields by hand, and manually create a new survey object with the appropriate observer configuration and data files. For more about survey objects, see Working with Surveys During Analysis in Chapter 7.

Output from DHT Analyses

The DHT engine produces the following output:

- a summary of results in the **Analysis Browser**. For general information about the Analysis Browser, see the section Introduction to the Analysis Browser in Chapter 7. There are many results statistics available, and you can select which ones are shown independently for each analysis set using the Column Manager (see Column Manager Dialog in the Program Reference).
- a detailed listing of results in the **Results** tab of the **Analysis Details** window. These are described in the following section, MRDS Results Details Listing.
- a log of the analysis, highlighting any possible problems, in the **Log** tab of the **Analysis Details** window. For information about troubleshooting problems, see Chapter 12 - Troubleshooting.

DHT Results Details Listing

When an analysis has run, information is available in the **Results** tab of the **Analysis Details** window. The results at present consist simply of a point estimate of abundance in the surveyed region.

Exporting DHT Results

The methods of exporting results from the Analysis Browser and results pages of the Analysis Details to other programs are the same as those for CDS analyses, as documented in Exporting CDS Results in Chapter 8. For example you can copy the results details text by choosing **Analysis Results | Copy Results to Clipboard.**

Miscellaneous DHT Analysis Topics

Clusters of Objects in DHT

If a cluster size field exists in the Observation layer, it becomes the numerator of the estimator shown at the beginning of this appendix.

If the objects detected are clusters of individuals, you tell Distance this in the same way as for CDS analyses, in the Setup Project Wizard (see Survey Methods Wizard Page in the Program Reference).

Unlike CDS analysis, the Model Definition does not offer any regression methods for dealing with size bias. If you suspect size bias is a potential problem, the appropriate way to deal with it in an MRDS analysis is to include cluster size (or some transformation of cluster size) as a covariate in the detection function model(s).

The cluster size field is one of the fields with a fixed name in detection function formulae in DHT (see Translating Distance Fields into DS and MR Covariates) – in formulae you should use the name `size` regardless of the actual field name.

Stratification and Post-stratification in DHT

At the moment, the MRDS engine only accommodates one level of stratification, and this stratification is assumed to be geographic. There is no allowance for weighting responses among strata in any fashion.

Running the DHT Analysis Engine from Outside Distance

The DHT engine is not yet implemented as a library in the free statistical software R. During this development phase, the DHT engine exists as a series of R scripts; these are not readily disseminated to the user community. In future, when DHT matures, we will distribute the code as an R library, and at that time the comments below will be applicable.

When you run a DHT analysis from Distance, Distance creates files containing a set of R commands and the appropriate input data, calls the R software, waits for the results and then reads them back in. For more about R, see R Statistical Software in Chapter 7 of the Users Guide.

Some users may wish to run the engine from outside the Distance interface – either from within the R GUI interface or from another program. For example, you may want to automate the running of analyses for simulations or bootstrapping.

To see the format of the input and data files produced by Distance, try running a DSM analysis in debug mode. To set debug mode on, choose **Tools | Preferences**, **Analysis** tab, and tick **Debug Mode**. When you run analyses in debug mode, the input and data files are created, but the analysis is not run. The Log tab displays the location of the files – these are created in a directory with a name "dst" followed by up to 4 numbers, located within the Windows temporary directory. The file "in.r" contains the commands. Data are located in the files "ddf.dat.r" (used in the detection function modelling), "region.dat.r", "sample.dat.r" and "obs.dat.r" (used in estimating density given a fitted detection function). You can use these files as templates for creating your own command and data files.

To run the analysis from within the R GUI (Graphical User Interface), you can cut and paste the commands from the file in.r. To run the analysis from another program, you can call R in batch mode – this is achieved by calling the program RCmd.exe, which is located within the /bin subdirectory of your R installation. For more details, see the R for Windows FAQ (in R, type `help.start()` and when a browser window opens, click on the FAQ for Windows port). For an example of its use, see the Log tab of any DSM analysis you have run that was not in debug mode – you should see a line of the form:

```
Starting engine with the following command:
C:\PROGRA~1\R\rw1091\bin\Rcmd.exe BATCH C:\temp\dst90474\in.r
C:\temp\dst90474\log.r
```

Users familiar with R may wish to work inside the R GUI. The DHT engine will be contained in the library DHT. To load the library from within R GUI, type

```
library(dht)
```

All the functions in the dsm library will be documented. You will be able to open a copy of the help files from within Distance by choosing **Help | Online Manuals | DHT Engine R Help (html)**.

The use of these libraries in operating systems other than Windows is not supported (but may well work – let us know!)

Checking Which Version of the DHT Engine is Being Used

The DHT engine is implemented as a library (called "DHT") in the statistical software R. From time to time, we may issue updated versions of the library, for example in response to reported problems. Therefore, before downloading a new version or reporting a problem you may want to check which version of the library is currently in use. To do this, re-run an analysis that uses the DHT Engine (such as one from the Williams sample project) and look in the Log tab for the line

```
> library(dht)
```

After it, you should see a line which looks something like the following:

The previous topic describes how to update to a newer version of the DHT Engine, if one is available.

When reporting results, you may want to cite the exact version (i.e., build number) of the library that used in the analysis. This is stored in the Log tab, as outlined above.

Bibliography

This section contains a list of references cited in the Users Guide. Much more complete lists of works related to distance sampling are in Buckland et al. (2001, 2004).

- Borchers, D.L., S.T. Buckland and W. Zucchini. 2002. *Estimating Animal Abundance: Closed Populations.* Springer Verlag.
- Borchers, D.L., S.T. Buckland, P.W. Goedhart, E.D. Clark and S.L Hedley. 1998a. Horvitz-Thompson estimators for double-platform line transect surveys. *Biometrics* **54**: 1221-37.
- Borchers, D.L., W. Zucchini and R.M. Fewster. 1998b. Mark-recapture models for line transect surveys. *Biometrics* **54**: 1207-1220.
- Buckland, S.T., D.R. Anderson, K.P. Burnham and J.L. Laake. 1993. *Distance Sampling: Estimating Abundance of Biological Populations.* Chapman and Hall, London, reprinted 1999 by RUWPA, University of St. Andrews, Scotland.
- Buckland, S.T., D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers and L. Thomas. 2001. *Introduction to Distance Sampling.* Oxford University Press, London.
- Buckland, S.T., D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers and L. Thomas. (eds.) 2004. *Advanced Distance Sampling.* Oxford University Press, London.
- Buckland, S.T., K.P. Burnham and N.H. Augustin. 1997. Model selection: an integral part of inference. *Biometrics* **53**: 603-618.
- Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd edition Springer-Verlag, New York.
- Gibbons, J.D. 1971. Nonparametric Statistical Inference. McGraw-Hill, New York.
- Hedley, S.L., S.T. Buckland and D.L. Borchers. 2004. Spatial distance sampling models. Pages 48-70 in Buckland, S.T., D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers and L. Thomas. (eds.) 2004. *Advanced Distance Sampling.* Oxford University Press, London.
- Horvitz, D.G., and D.J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**: 663-685.
- Innes, S., Heide-Jørgensen, M.P., Laake, J.L., Laidre, K.L., Cleator, H.J., Richard, P. and Stewart, R.E.A. (2002) Surveys of

belugas and narwals in the Canadian High Arctic in 1996. *NAMMCO Scientific Publications* **4**: 169-190.

- Laake, J.L. and D.L. Borchers. 2004. Methods for incomplete detection at distance zero. Pages 108-189 in Buckland, S.T., D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers and L. Thomas. (eds.) 2004. *Advanced Distance Sampling.* Oxford University Press, London.
- Marques, F.F.C. 2001. *Estimating wildlife distribution and abundance from line transect surveys conducted from platforms of opportunity.* PhD Dissertation, University of St Andrews, Scotland.
- Marques, F.F.C. and S.T. Buckland. 2003. Incorporating covariates into standard line transect analysis. *Biometrics* 59: 924-935.
- Marques , F.F.C. and S.T. Buckland and D.L. Borchers. 2004. Covariate models for the detection function. In Buckland, S.T., D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers and L. Thomas. (eds.) In prep. *Advanced Distance Sampling.* Oxford University Press, London.
- Marques, T. A., Thomas, L., Fancy, S. G. , and S. T. Buckland. 2007. Improving estimates of bird density using multiple covariate distance sampling. The Auk. 124:1229-1243.
- Otto, M.C. and K.H. Pollock. 1990. Size bias in line transect sampling: a field test. *Biometrics* **46**: 239-45.
- Strindberg, S. 2001. *Optimized Automated Survey Design in Wildlife Population Assessment.* PhD Dissertation, University of St Andrews, Scotland.
- Strindberg, S. and S.T. Buckland. 2004. Zigzag Survey Designs in Line Transect Sampling. *Journal of Agricultural, Biological & Environmental Statistics* 9: 443 - 461.
- Strindberg , S., S.T. Buckland. and L. Thomas. 2004. Survey design and Geographic Information Systems. In Buckland, S.T., D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers and L. Thomas. (eds.) *Advanced Distance Sampling.* Oxford University Press, London.
- Thomas, L., R. Williams and D. Sandilands. 2007. Designing line transect surveys for complex survey regions. *Journal of Cetacean Research and Management*, in press.
- Thompson, S.K. *Sampling*. (2nd edition) 2002. John Wiley and Sons, New York.

Glossary of Terms

adjusted angle zigzag

Survey design class that superimposes a continuous zigzag sampler whose angle is continuously adjusted by survey region height.

AIC

Akaike's Information Criterion (AIC) is used in model selection and puts this process into a function minimization framework. It is based on the Kullback-Leibler "distance" between two distributions.

For more about AIC and model selection, see Burnham and Anderson (2002).

AICc

Version of AIC corrected for small sample size.

For more information, see Burnham and Anderson (2002).

analysis engine

A component within Distance that runs analyses and produces results. Different analysis engines have different capabilities. Currently, there are three analysis engines, one for conventional distance sampling (CDS), one for multiple covariate distance sampling (MCDS), and one for mark-recapture distance sampling (MRDS).

API

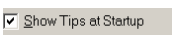
Abbreviation for Application Programming Interface – an interface that allows a piece of software to be instructed to perform tasks from within a separate software package.

CDS

See conventional distance sampling

checkbox

A box in the graphical user interface of Distance that you click on to select. A selected checkbox displays a tick. Example:



conventional distance sampling

A subset of distance sampling methods, where probability of detection on the point or line is assumed to be 1, and the only covariate in the detection function is distance. For more details, see Buckland et al. (1993, 2001).

covariate

A variable that you can use to model the detection function. Perpendicular or radial distance is always used as a covariate, but in the Multiple Covariate Distance Sampling (MCDS) engine, you can include other covariates, such as cluster size, sex, platform of observation, habitat, etc.

coverage probability

The coverage (or inclusion) probability of at an arbitrary location within the survey region is the probability of it falling within the sampled portion of the survey region.

data file

This file, always called DistData.mdb, contains information about how the data is stored, and may contain some or all of the data itself. It is stored in the data folder.

data folder

A folder (directory) containing survey data and related information about the survey effort, study area boundaries, etc. Data folder names always have the same beginning as the associated project file, but end in .dat - e.g., Ducknest.dat. The data folder contains the data file, and one or more other files.

detection function

A function, denoted $g(x)$, that described the probability of detecting an object (individual or cluster) given that it is at distance x from the transect line or point. In Distance, the detection function is modeled using the key + series adjustment framework described in Buckland et al. (1993, 2001).

densification

A line that is straight in one coordinate system will not necessarily be straight when viewed in a different system. For example, the equator is not a straight line on many maps. So, when projecting from one coordinate system to another, a straight line must be broken into a series of smaller straight lines so that it stays in approximately the same place in the projected coordinate system. This process of adding vertices to a line when projecting it is called densification.

design axis

User-defined line, superimposed on the survey region, that is used to orient the samplers in zigzag designs.

dialog

A type of window in the graphical user interface of Distance. Dialog windows are *modal* – that is you cannot access any other windows in Distance until you

close the dialog. Only one dialog can be open at once. Examples include the properties windows (e.g., Model Definition Properties), and the Open Project dialog.

distance project

Where all of the information about one study area is stored. A project is made up of a project file (which ends in .dst) and a data folder (ends in .dat).

distance sampling

A group of related survey methods for estimating the density and/or abundance of wildlife populations.

double observer

A survey protocol where two (semi-) independent observer teams perform a distance sampling survey, and duplicate detections are identified. Under this protocol, more advanced analysis methods (Mark Recapture Distance Sampling) can be used where it is possible to relax the assumption of standard methods that all animals at zero distance are seen.

For more information, see Laake and Borchers (2004). For more about how to set up a double observer dataset in Distance, see the Users Guide chapter on Mark Recapture Distance Sampling.

equal angle zigzag

Survey design class that superimposes a continuous zigzag sampler of fixed angle on the survey region.

equal spaced zigzag

Survey design class that superimposes a continuous zigzag sampler that passes through equally spaced points on opposite sides of the survey region boundary.

external data files

Files that contain information about the survey data in a project, other than the main data file DistData.mdb.

factor

Name given to a covariate that is divided into distinct classes. Examples include sex (male / female), observer, etc.

f(0)

The value of the probability density function of observed distances, evaluated at 0 distance.

geographic coordinates

A measurement of a location on the earth's surface expressed in degrees of latitude and longitude.

geographic coordinate system

A reference system used to locate points expressed in degrees of latitude and longitude on the earth's surface. Defined by a spheroid of reference, a datum, one or more standard parallels, a central meridian, and possible shifts in the x- and y-directions to locate x,y positions of point, line, and area features

GIS

Geographic Information System - a piece of software that can work with geographic data.

Horvitz-Thompson

Unbiased estimator of abundance, given by

$$\hat{N}_{surv} = \sum_{i=1}^{n} \frac{1}{p_i}$$

where N_{surv} is the number of animals in the surveyed area (i.e., the strip or circle actually surveyed), n is the number of animals seen, and p_i is the probability of observing the ith animal, given that it is in the surveyed region.

Given this estimate, assuming equal probability of coverage, an estimate of the population abundance N is given by

$$\hat{N} = \frac{A}{a} N_{surv}$$

where A is the area of the survey region, and a is the surveyed area.

If coverage probability is not equal, population abundance can be estimated by

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{p_i q_i}$$

where q_i is the probability that surveyed area covers the ith animal, given its location. q_i is dictated by the survey design.

For more information, see Horvitz and Thompson 1952, Borchers et al. 2002, Thompson 2002, Buckland et al. In prep

Horvitz-Thompson-like

Term used to describe a Horvitz-Thompson estimator, where the probability of observing the animal is estimated, rather than known:

$$\hat{N}_{surv} = \sum_{i=1}^{n} \frac{1}{\hat{p}_i}$$

This estimator is biased, although the bias is usually not large if the p_is are not small.

See the entry for Horvitz-Thompson estimator for notation, and generalization to the case where coverage probability is not equal.

For more information, see Horvitz and Thompson 1952, Borchers et al. 2002, Thompson 2002, Buckland et al. In prep

inclusion probability

see Coverage Probabiltiy

mark recapture distance sampling

A type of distance sampling used when probability of detection on the transect line or point is less than 1. Multiple observers survey independently (or semi-independently), and duplicate detections are recorded. For more details, Chapter 6 of the Advanced Distance Sampling book.

MCDS

see multiple covariate distance sampling

MRDS

See mark recapture distance sampling

multiple covariate distance sampling

Analysis of distance sampling data where covariates in addition to distance are used to model the detection function. For more details, see Chapter 3 of the Advanced Distance Sampling Book.

multiplier

A quantity you can use when you know your estimates are proportional to the true abundance or density. If you know the constant of proportionality, you can use a multiplier to get unbiased estimates. An example would be if you know that g(0) is less than 1, but you have an independent estimate of g(0). You can then use the multiplier (and, if you have it, the multiplier SE and DF) to correct your estimates.

parallel random sampling

Survey design class that randomly distributes a number of parallel lines over the survey region.

probability of detection

The probability of recording an object (individual or cluster) in the surveyed area.

project file

A file containing the project settings, survey designs, analysis settings and results. Project files always end in .dst – e.g., Ducknest.dst. Double-clicking a project file opens it in Distance.

projected coordinates

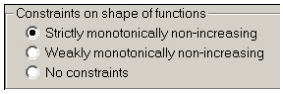
A measurement of locations on the earth's surface expressed in a two-dimensional system that locates features based on their distance from an origin (0,0) along two axes, a horizontal x-axis representing east–west and a vertical y-axis representing north–south. A map projection transforms latitude and longitude to x,y coordinates in a projected coordinate system.

projected coordinate system

A reference system used to measure horizontal and vertical distances on a map. In Distance, a projected coordinate system is defined by a map projection (which may include projection parameters such as shifts in the x or y direction) combined with a geographic coordinate system.

radio button

A round button that you click on to select. Radio buttons usually occur in a group, of which only one can be selected at once. An example is the constraints group in the Conventional Distance Sampling Model Definition properties window:



R folder

A folder (directory) containing the R object file (.RData) and image files generated by the R statistical software package. It is located within a project's data folder, and is created automatically the first time an analysis is run that uses R.

R software

According to the R web site (http://www.r-project.org), R is a language and environment for statistical computing and graphics. In the context of Distance, the mark-recapture distance sampling (MRDS) analysis engine is implemented as an R library. A working copy of R is therefore required before this engine can be run.

set

A collection of Analyses, Designs or Surveys that are displayed on the same browser page. You usually group items together that share some properties – for example, you could have two different Analyses Sets, one where you use truncation and one where you do not.

shapefile

A shapefile is a standard format for storing geographic information, invented by the GIS company ESRI. Each shapefile is actually 3 separate files: an .shp file, a .shx file and a .dbf file. (In addition, there may be other files such as .prj files.) Shapefiles are used to store geographic information in Distance.

simple random sampling

Survey design class that randomly distributes a fixed number of points over the survey region.

single observer

The standard survey protocol where a single team of observers perform a distance sampling survey. Under this protocol, it is necessary to assume that all animals at zero distance are detected. cf. double observer methods.

systematic grid sampling

Survey design class that randomly superimposes a systematic point grid of fixed dimensions and rotation onto the survey region.

systematic random sampling

Survey design class that randomly superimposes a systematic set of parallel lines onto the survey region.

systematic segmented grid sampling

Survey design class that randomly superimposes a systematic set of segmented parallel lines onto the survey region. Segments are placed using a grid of points.

systematic segmented trackline sampling

Survey design class that randomly superimposes a systematic set of segmented parallel lines onto the survey region. Segments are evenly spaced along a systematic series of parallel tracklines.

toolbar

The collection of buttons and menus at the top of a window in Distance. An example is the Analysis Components toolbar:

trackline

The transect line.

Index

A

About Distance 5
About Distance dialog 254
About the Users Guide 1
Acknowledgements 6
Adjustment terms
  Specifying in CDS and MCDS analysis 241
Algorithms
  MCDS engine 314
Analyis Results
  Output from CDS analyses 89
  Output from DSM analyses 154, 329
  Output from MRDS analyses 135
Analysis
  Stopping 163
Analysis Browser 71, 197
  CDS Results 98
  DSM Results 156
  Exporting CDS Results 98
  MRDS Results 136
Analysis Components 72, 253
Analysis Components Window 77, 253
Analysis Details 72
  CDS Results 90
  DSM Results 155, 329
  Exporting Results 99
  MCDS results 122
  MRDS Results 135
Analysis Details window 210
Analysis Details Windows 72
Analysis Engines
  About 80
  CDS Output 89
  DSM Output 154, 329
  MCDS output 122
  MRDS Output 135
  Running DSM from outside Distance 157, 330
  Running MCDS from the command line 279
  Running MRDS from outside Distance 140
Analysis Guidelines
  CDS 88
  DSM 152
  MCDS 120
  MRDS 134
Analysis in Distance
  Analyis Details window 210
  Analysis Browser 197
  Analysis components 72
  Analysis engines 80
  Introduction 71
  Preferences 181
Analysis Results
  Output from MCDS Analyses 122
Authors 6

B

Backing up projects 36
Bibliography 333
Binned Data
  CDS 100
Books
  Bibliography 333
  Distance sampling reference books 3
Bootstrap
  Overview in CDS 107
  Setting options in CDS and MCDS analysis 247
Bootstrap file
  MCDS engine file format 313
Bootstrap progress file
  MCDS engine file format 313
Bounds on parameters
  Setting in CDS and MCDS 244
  Specifying in DSM 159
  Specifying in MRDS 142

C

Calculating Probability of Detection 95
CDS *See* Conventional Distance Sampling
Citation for Distance 3
Cleaning the Windows Temp folder 83
Cluster Size tab
  CDS and MCDS 246
Clustered populations
  In CDS 102
Clusters
  About 102
  CDS 102
  Data Filter options 233
  Missing values 102
  Model Definition options 246
  Zero cluster size 102
Column Manager dialog 258
Compacting a project 39
Components that make up the Distance software 261
Constraints
  Setting in CDS and MCDS 244
Control
  Specifiying control parameters in MRDS Model Definition 252
Conventional Distance Sampling 87
  About 87
  Analysis guidelines 88

Limitations of engine 314
MCDS engine reference 279
Output from CDS analyses 89
Running from command line 279
Setting up a CDS Project 88
Conventions used in this documentation 2
Coordinate system, geographic 53
Copy
Data to other Programs 187
Corrupted project
Fixing 164
Covariates
Density Surface Modelling 145, 327
Factor vs. non-factor in MCDS 117
Factor vs. non-factor in MRDS 134
Mark Recapture Distance Sampling 127
Multiple Covariates Distance Sampling 115
Specifying in MCDS analysis 243
Coverage Probability 65
Cramér-von Mises Test 93
Creating a new project 34

D

Data
For CDS analysis 88
For DSM analysis 146, 329
For MCDS analysis 116
For MRDS analysis 129
Data Entry 46
Data Entry Wizard 170
Data Explorer 183
Data Fields 43
Data file reference 262
Data Filter
About analysis components 72
Interface 231
Data Filter Properties dialog 231
Data Format
GIS data 56
Data Import
Advanced 48
Getting started example 1 13
Getting started example 2 19
GIS data 57
Import Data Wizard 172
Importing one file per data layer 50
Introduction 46
Linking to data from other databases 61
Non geographical data 46
Streamlining import of data from one flat file 49
Troubleshooting 176
Data in Distance 41
Data structure 41
Getting data into Distance 45
How Distance stores data internally 262
Data Layer Properties 256
Data Layers
About 41
List of layer types 42
Data Selection 232
Data structure
About 41
Changing 45
Database API 261
Density Surface Modelling
About model formulae 151
Analysis guidelines 152
Checking the version number 158, 331
Fine-tuning an MRDS analysis 159
Installing an updated version of the engine 158
Introduction 145, 327
Output from DSM analyses 154, 329
Running from outside Distance 157, 330
Setting up a DSM Project 146, 329
Design Browser 193
Design Classes 64
Design Details window 201
Design Properties dialog 215
Detection function
Constraints
Setting in CDS and MCDS 244
Specifiying in CDS and MCDS analyses 240
Specifiying in MRDS analyses 251
Detection Function
About Formulae 95
Multiple detection functions 111
Detection Function tab
CDS and MCDS 240
MRDS 251
Development Team 6
Diagnostic output
Specifying in CDS and MCDS analysis 245
Specifying in MRDS analysis 252
Distance
Data file reference 262
Distance
About 5
Authors 6
Citation 3
Components that make up the software 261
Data structure 41
Database API 261
History 7
Inner workings 261
New features 8
Sponsors 5
Use Agreement 5
Web site 4
Distance projects 33
Distance sampling
Reference books 3
Distance-sampling email list 3
Double observer configuration
Analysis using CDS and MCDS engine 125
Analysis using MRDS engine 127
Project setup 166
Double observer methods 127
DS model
About, in MRDS Engine 131

DS Model
  Specifying in Model Definition 251
DS model formulae
  About in MRDS Engine 131
DSM *See* Density Surface Modelling, *See* Density Surface Modelling

E

Edge Effects 65
Email list 3
Equal Spacing Zigzag Design
  First and Last Line Placement 67
Errors 161
  In CDS and MCDS engine 162
  Internal errors in CDS and MCDS analysis engines 162
  Internal errors in the interface 161
  Known problems 161
Estimate tab
  CDS and MCDS 237
  MRDS 250
Estimating density from a subset of the data in MRDS 139
Estimating the Detection Function at Multiple Levels in MCDS 118
Example
  Mexico survey design 26
  More Complex Data Import 19
  Sample Projects 31
  St Andrews Bay survey design 21
  Survey Design 21
  Using Distance to Analyse Simple Data 13
Exporting Data 187
Exporting Projects 38

F

Factor and non-factor covariates
  In MCDS 117
  In MRDS 134
Factor covariates
  Specifiying in MRDS Model Definition 252
Field names
  Valid names 276
Fine-tuning a DSM analysis 159
Fine-tuning an MRDS analysis 142
Formulae
  About in DSM Engine 151
  About in MRDS Engine 131
  About, in CDS Engine 95

G

g(0)<1
  CDS and MCDS via Multipliers 108
  MRDS 127
Geographic coordinate system 53
Geographic Data 52
Geographic data in Distance 52
Geographic projection 53
Getting Started
  Analysis 1 13
  Analysis 2 19
  Objective 13
  Sample Projects 31
  Survey Design 21
GIS data
  Format 56
GIS data
  Preferences 180
GIS Data
  Troubleshooting 163
GIS Data
  About 52
  Copy and paste from Clipboard 57
  Importing 57
  Viewing and Manipulating 52
Goodness of fit
  Chi-square in MCDS 124
  Cremér-von Mises test 93
  Kolmogorov-Smirnov test 93
  Specifying in CDS and MCDS analysis 245
  Specifying in MRDS analysis 252
Grouped Data
  CDS 100

H

History of Distance 7

I

Images produced by R 84
Import data
  Non geographical data 46
Import Data
  Covariate data example 19
  Getting started example 1 13
  Getting started example 2 19
Import Data Wizard 172
  About 172
  Troubleshooting 176
Importing exisisting GIS Data 57
Importing from previous versions of Distance 39
Inferences just on Covered Region
  CDS and MCDS 112
  MRDS 140
Inside Distance 261
Internal errors 161
Interval Data
  CDS 100
Intervals 233
Iterations
  Showing and controlling maximum number in MRDS 142

K

Key function

Specifying in CDS and MCDS analysis 240
Known problems 161
Kolmogorov-Smirnov test 93

L

Limitations
MCDS engine 314
Linking to external data 61
Locking the data sheet 82
Log file
MCDS engine file format 311

M

Manual selection
Of adjustment terms in CDS and MCDS analysis 241
Map Browser 192
Map Properties 258
Map window 199
Maps
Map Browser 192
Map window 199
Mark Recapture Distance Sampling
Analysis guidelines 134
Checking the version number 141
Defining MRDS models 130
Factor vs. non-factor covariates 134
Fine-tuning an MRDS analysis 142
Installing an updated version of the engine 141
Introduction 127
Output from MRDS analyses 135
Running from outside Distance 140
Setting up an MRDS Project 129
Troubleshooting problems 162
Using a previously fitted detection function 139
Maximum iterations
Setting in MRDS analysis 142
Maximum Observations, Samples etc *See* Limitations
MCDS *See* Multiple Covariates Distance Sampling
MCDS engine
Changes 324
MCDS engine fitting algorithms 314
MCDS engine reference 279
Misc. tab
CDS and MCDS 249
Missing cluster sizes
CDS 102
Missing covariate values in MCDS 124
Missing data
CDS 101
MCDS 124
Missing cluster sizes in CDS 102
Model Averaging
CDS and MCDS 111
MRDS 138
Model Definition
Interface 236
Model Definition Properties dialog 236
MR model
About, in MRDS Engine 131
MR Model
Specifying in Model Definition 252
MR model formulae
About in MRDS Engine 131
MRDS *See* Mark Recapture Distance Sampling
Multi-model inference 111
Multiple Covariate Distance Sampling
Setting up an MCDS Project 116
Multiple Covariates Distance Sampling 115
Analysis guidelines 120
Defining MCDS models 116
Estimating the detection function at multiple levels 118
Factor vs. non-factor covariates 117
Introduction 115
Limitations of engine 314
MCDS engine reference 279
Output from MCDS analyses 122
Scaling of distances for adjustment terms 119
Multipliers
CDS and MCDS 108
MRDS 138
Multipliers tab
CDS and MCDS 247
Multi-species Study
Estimating detection function for rarer species 108

N

New Features 8
In Distance 3.5 10
In Distance 4.0 8
In Distance 4.1 8
In Distance 5.0 8
In Future Versions 12
New Project
Creating a New Project 34
Setup Project Wizard 165
Non-convex Regions
Zigzag Sampling in 67

O

Opening a project 36
Output file
MCDS engine file format 310

P

Parametric Indexing 97
Plot file
MCDS engine file format 312
Plots
Exporting CDS plots 99
Exporting CDS plots to R 99
Exporting MCDS plots 124
Exporting MCDS plots to R 124
In CDS Results 90

In MCDS Results 122
Produced by R 84
Qq-plots 92
Saving to file in CDS and MCDS analysis 245
Post-stratification
CDS 103
MCDS 125
MRDS 137
Preferences 178
Probability of Coverage 65
Probability of Detection
Calculating 95
Problems
CDS and MCDS engines 162
GIS 163
Known problems 161
MRDS engine 162
Recovering from unexpected program exit 164
Reporting 4
Troubleshooting 161
With the Analysis Engines 162
Program Reference 165
Project
About 33
Archiving 38
Backing up 36
Compacting 39
Creating 34
Editing 39
Exporting 38
Importing 39
Opening 36
Sample Projects 31
Saving 36
Template 35
Transporting 38
Viewing 39
Project Browser 183
Project Properties 177
Projection Parameters dialog 255
Projection, geographic 53

Q

Qq-plots
CDS 92
Qq-Plots
Specifying in CDS and MCDS analysis 245

R

R 83
About the link between R and Distance 83
Density Surface Modelling 145, 327
dsm library 157, 330
Folder 84
Images, about 84
Installing and Configuring 84
Mark Recapture Distance Sampling 127
mrds library 140
Preferences 181
R Image Properties dialog 259, 260
Supported versions 83
Updating the Version that Distance Uses 84
R Folder 84
R statistical software 83
Random number generation algorithms 278
Recovering from unexpected program exit 164
References 333
Reporting problems 4
Reserved field names 276
Results
CDS analysis 89
DSM analysis 154, 329
Exporting CDS output 98
MCDS analysis 122
MRDS analysis 135
Running Analyses 81
Running DSM engine outside the Distance interface 157, 330
Running MCDS engine outside the Distance interface 279
Running MRDS engine outside the Distance interface 140
Running the MCDS engine as a stand-alone program 279
Saving CDS and MCDS results to file 100

S

Sample definition
In CDS Analysis 112
In MRDS Analysis 139
Sample Projects 31
Saving a project 36
Saving CDS Analysis Results
Cutting and pasting 99
Saving to file 100
Saving MCDS Analysis Results
Cutting and pasting 124
Scaling of distances for adjustment terms in MCDS 119
Setup Project Wizard 165
Single observer configuration
Analysis using MRDS engine 143
Project setup 166
Single transect
CDS and MCDS 113
Smearing 249
Sponsors 5
Starting values
Specifying in CDS and MCDS analysis 241
Specifying in DSM analysis 159
Specifying in MRDS analysis 142
Stats file
MCDS engine file format 311
Stopping an analysis 163
Stratification
CDS 103
DSM 157, 330

MCDS 125
MRDS 137
Suggestions
Sending suggestions 4
Survey Browser 195
Survey Design in Distance
Design Browser 193
Design Details window 201
Design Properties dialog 215
Introduction 63
Preferences 180
Setting up a new project 68
Survey Browser 195
Survey Details window 209
Survey Details window 209
Survey Properties dialog 230
Surveys
Working with surveys during analysis 78

T

Temp folder
Cleaning 83
Template, using a project as 35
Troubleshooting 161
Truncation 235

U

Units 236
Unknown Study Area Size 112
Use Agreement 5
Using a previously fitted detection function to estimate density in MRDS 139

V

Valid field names 276
Variance Estimation
CDS 107
MRDS 137
Variance tab
CDS and MCDS 247
MRDS 252, 253
Version
Distance 254
MRDS Engine 141, 158, 331

W

Warnings
In CDS and MCDS engine 162
Web site 4
Welcome to the Users Guide 1
What is Distance? 5
Which geographic projection? 55
Wizards
Data Entry Wizard 170
Setup Project Wizard 165

Z

Zigzag Sampling Designs
First and Last Line Placement 67
Introduction 66
Non-convex regions 67
zip files
Exporting projects as 38